A Generalization Theory for Zero-Shot Prediction

Ronak Mehta [1] **Zaid Harchaoui** [1]

Abstract

A modern paradigm for generalization in machine learning and AI consists of pre-training a task-agnostic foundation model, generally obtained using self-supervised and multimodal contrastive learning. The resulting representations can be used for prediction on a downstream task for which no labeled data is available. We present a theoretical framework to better understand this approach, called zero-shot prediction. We identify the target quantities that zero-shot prediction aims to learn, or learns in passing, and the key conditional independence relationships that enable its generalization ability.

1. Introduction

In 2021, OpenAI shocked the world by improving the zero-shot classification accuracy on ImageNet from 11.5% to 76.2% via the CLIP series of models (Radford et al., 2021). This event redefined the goal of zero-shot prediction from producing models that generalized to *unseen classes* to those that generalized to *unseen tasks* entirely. Two fundamental drivers of CLIP's success were 1) the use of natural language as a medium for representing arbitrary classes (as in the previous state-of-the-art Visual N-grams (Li et al., 2017)), and 2) a massive, yet carefully designed pre-training set which significantly impacted downstream performance (Radford et al., 2021; Fang et al., 2023; Xu et al., 2024). Despite the remarkable success of these foundation model-based pipelines (Bommasani et al., 2022), there are unique components of zero-shot prediction that warrant investigation from a theoretical point of view.

To clarify these gaps, we contrast zero-shot prediction (ZSP) with the related setting of few-shot learning (FSL). Let $x \in \mathcal{X}$ denote an input (often an image) that accompanies a discrete value $y \in \mathcal{Y}$ (often a class label). Common to both ZSP and FSL is a pre-training procedure in which a large unlabeled dataset $x_1, \ldots, x_N \in \mathcal{X}$ is used to produce an *encoder* $\boldsymbol{\alpha} : \mathcal{X} \to \mathbb{R}^d$. The *embedding* $\boldsymbol{\alpha}(x)$ is thought to contain information that is relevant for predicting y from x. Pre-training typically occurs through the process of self-supervised learning (SSL), using a *pretext task* that can be solved with only instances of x (e.g. filling in a blank image patch). In FSL, the user may then access a labeled dataset $(x_1^{\text{lab}}, y_1^{\text{lab}}), \ldots, (x_n^{\text{lab}}, y_n^{\text{lab}})$ from which a predictor can be trained inexpensively. This often takes the form of a linear classifier $x \mapsto \mathbf{W}\boldsymbol{\alpha}(x) + b$ for $\mathbf{W} \in \mathbb{R}^{|\mathcal{Y}| \times d}$ and $b \in \mathbb{R}^{|\mathcal{Y}|}$. Where ZSP departs from FSL is the additional challenge of being given *no directly labeled training data*.

At first glance, ZSP seems impossible. Yet, the ingenuity of practitioners has yielded the following solution; if 1) each pre-training example x_i is paired with another "view" $z_i \in \mathcal{Z}$ (e.g. a caption in natural language) and 2) if each label $y \in \mathcal{Y}$ can intelligently be embedded into $\mathcal{Z}$, then the relationship between each x_i and z_i could provide the means to perform prediction. Concretely, one learns a complementary encoder $\boldsymbol{\beta} : \mathcal{Z} \to \mathbb{R}^d$ during pre-training and designs *prompts* z_k^y for $y \in \mathcal{Y}$ and $k = 1, \ldots, m$. Then,

$$x \mapsto \arg\max_{y \in \mathcal{Y}} \frac{1}{m} \sum_{k=1}^{m} \langle \boldsymbol{\alpha}(x), \boldsymbol{\beta}(z_k^y) \rangle \qquad (1)$$

is employed for prediction. An example of a prompt is the template text "photo of a ___.", where the blank can be filled by the textual representation of the class (e.g. "cat" or "dog"). The ZSP pipeline, from pre-training to prompt selection, is clearly a wild departure from what is explained by statistical learning theory. Moreover, while some components of these systems have been studied in the context of FSL (such as the reasons why various pre-training objectives result in encoders that provably accelerate learning), unique aspects of ZSP, such as the role of prompting and the cost of "translating" modalities, have not yet received theoretical treatment. Herein lies our question.

Through what decomposition of downstream task performance can we compare zero-shot prediction to the direct supervised learner, with a transparent dependence on the 1) pre-training distribution, 2) evaluation distribution, and 3) prompting strategy?

[1]Department of Statistics, University of Washington, Seattle. Correspondence to: Ronak Mehta <ronakdm@uw.edu>.

Contributions. In Sec. 2, we present a learning theoretic framework for the pre-training/evaluation/prompting data and propose two expressions for the population counterpart of (1). These expressions, while equivalent at the population level, reflect two classes of learning methods which we call the "conditional mean" and the "information density" approaches. In Sec. 3, we prove a generic decomposition of the prediction error on the downstream task, which furnishes three components: *prompt bias* measures the compatibility of the prompt strategy with the pre-training and evaluation distributions, *residual dependence* measures the information-theoretic cost of using one modality to make predictions on another, and *estimation error* quantifies the effect of the finite number of pre-training examples and prompts. The estimation error decomposes further depending on whether the conditional mean or information density approaches are taken. To provide insight and demonstrate the usefulness of the decomposition, we analyze the performance of nonparametric regression methods for each approach by way of finite-sample bounds in high probability. Our framework arms practitioners with a means to imbue existing SSL-to-ZSP pipelines with theoretical guarantees, depending on the approach with which they best align. In Sec. 4, we illustrate our theoretical claims by empirically evaluating prompt bias and residual dependence on zero-shot prediction tasks with simulated and image data.

Related Work. One can argue that precursors to both FSL and ZSP in machine learning can be found in the literature of meta-learning, or "learning to learn" (Thrun and Pratt, 1998; Andrychowicz et al., 2016; Finn et al., 2017). There, the downstream evaluation tasks are given to the user upfront, so that pre-training an encoder and training a predictor for all of the evaluation tasks can be performed in one step. On the other hand, FSL and ZSP both involve fully task-agnostic pre-training phases. Seminal work in computer vision on matching words and pictures is also worth mentioning (Barnard et al., 2003; Forsyth et al., 2009).

Two complementary bodies of work studied phenomena common to FSL and ZSP. The first considers which properties of learned encoders can provably improve downstream performance (Wang and Isola, 2020; HaoChen et al., 2021; Atzmon et al., 2020; Wang and Jordan, 2024; Du and Xiang, 2024). The other is dedicated to explaining how otherwise mysterious SSL objectives achieve these properties (Wen and Li, 2021; Li et al., 2021; Pokle et al., 2022; Kiani et al., 2022; Johnson et al., 2023; Shwartz-Ziv et al., 2023). In particular, Balestriero and LeCun (2022) and Tan et al. (2024) relate various SSL objectives to spectral clustering. One FSL-specific line of work studies when linear mappings of pre-trained encoders can achieve optimal downstream performance (Saunshi et al., 2019; HaoChen et al., 2021; Tosh et al., 2021; Lee et al., 2021).

While informative representations are essential, the core of ZSP is the remarkable ability of models to make predictions without *any* task-specific data, a challenge even for the perfect encoder. For context, we avoid the historical term "zero-shot learning" (Larochelle et al., 2008; Akata et al., 2015), which refers to a setting in which pre-training data is not only *labeled*, but contains metadata-based features associated with each class. In general, this only handles unseen classes, and only if the same features are observed at inference time. To our knowledge, the only work studying ZSP based on self-supervised pre-training is Chen et al. (2024). In particular, Chen et al. (2024, Theorem 4.2 and Corollary 5.1) provides bounds on the top-k accuracy of ZSP for CLIP-based encoders. However, the bound *increases* with the batch size, may not decay to zero even if the pre-training loss is fully optimized and upstream and downstream data distributions are the same, and does not seem to explicitly depend on the prompt quality. The independent and concurrent work of Oko et al. (2025) develops a statistical analysis based on sufficiency notions, with the aim of capturing the predictive performance in the downstream task. Their work is complementary to ours, in that they determine the distributional parameter learned by the CLIP objective, but also assume that the prompting strategy and downstream data distribution are "idealized", in that the prompt bias and residual dependence quantities alluded to in the contributions are zero.

On the applied side, we are inspired by the number of works that use diverse, class-specific prompts generated using large language models (LLMs) for enhancing ZSP performance (Pratt et al., 2023; Yang et al., 2023; Maniparambil et al., 2023) and interpretability (Menon and Vondrick, 2023; Esfandiarpoor et al., 2024). While these empirical methods, such as the customized prompts via language models method (CuPL, Pratt et al. (2023)), are often designed using intuition from human understanding of natural language, we aim to offer a theoretical explanation for their success from a statistical learning theory and probabilistic graphical modeling perspective. Despite this particular application of LLMs, we also acknowledge that "prompting" in ZSP has a different meaning than in the growing field of prompt engineering, in which inputs are designed for large language models (Pryzant et al., 2023; Wang et al., 2024; Guo et al., 2024; Sclar et al., 2024).

2. Theoretical Framework

We introduce the mathematical objects that connect the empirically-motivated predictor (1) to its theoretical counterpart analyzed in Sec. 3. For the reader's convenience, a global notation table is provided in Appx. A.
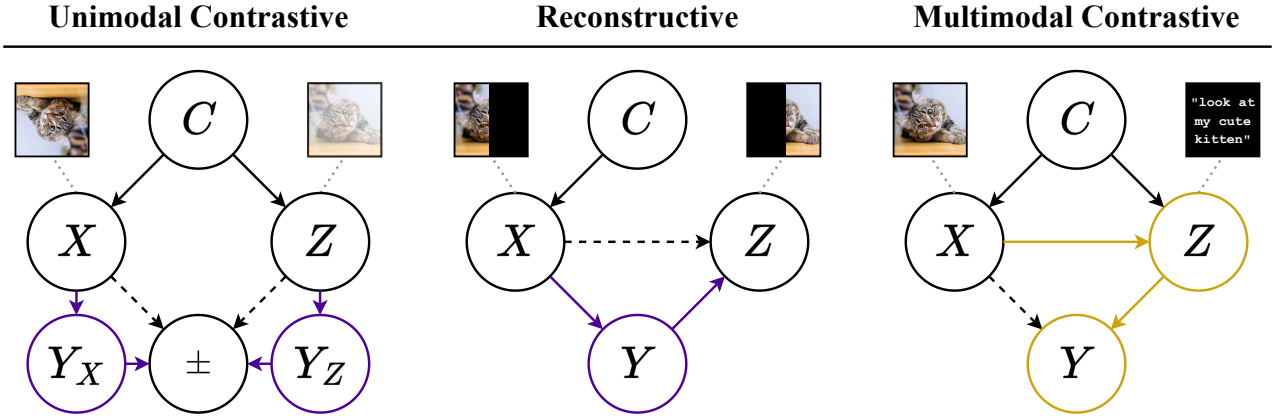
Unimodal Contrastive   **Reconstructive**   **Multimodal Contrastive**

Figure 1. **Graphical Models of Prediction Paths.** Each directed graphical model corresponds to the data types and dependence structures for various SSL pre-training approaches. The variable C represents an unobserved context that determines the observed data-generating distribution. Dotted lines indicate the possibility of presence or absence of the arrow. Methods **compatible with ZSP** may learn the relationship between X and Z directly, whereas the relationship between Z and Y is learned via prompting. Methods that are **compatible with FSL** learn the label Y as a latent variable in the process of solving the pretext task.

Prediction Setups. Consider random variables X, Y, and Z observed in $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$, respectively. We interpret $\mathcal{X}$ as the space of images, $\mathcal{Y}$ as the (not necessarily discrete) space of labels, and $\mathcal{Z}$ as the space of text captions.

Consider a probability measure $P_{X,Y}$ on $\mathcal{X} \times \mathcal{Y}$, called the *evaluation distribution*. We specify a collection of downstream tasks, with which we may evaluate predictors on data drawn from $P_{X,Y}$ (e.g. CIFAR-10). Consider a function $r : \mathcal{Y} \to \mathbb{R}$, and the least squares prediction problem

$$\min_{\eta:\mathcal{X}\to\mathbb{R}} \mathbb{E}_{P_{X,Y}} \left[(\eta(X) - r(Y))^2 \right] \qquad (2)$$

The function r serves only to handle multiple task formats such as regression ($r(\boldsymbol{y}) = \boldsymbol{y}$) or binary classification ($r(\boldsymbol{y}) = \mathbb{1}\{\boldsymbol{y} = 1\}$) in a unified manner. We discuss formulations of multi-class classification and structured prediction in Sec. 3. The optimizer of (2) over $\eta \in \mathbf{L}^2(P_X)$[1], or all measurable, square-integrable functions on $\mathcal{X}$, is

$$\eta_\star(\boldsymbol{x}) := \mathbb{E}_{P_{Y,X}} \left[r(Y)|X \right] (\boldsymbol{x}). \qquad (3)$$

We will call this the *direct predictor* throughout this paper, which will contrast our viewpoint of ZSP as an indirect, multi-stage prediction procedure. Indeed, the prompting step in (1) resembles an empirical average of draws from a probability distribution on $\mathcal{Z}$ based on the class label $Y = \boldsymbol{y}$ (especially when considering the LLM-based generation methods mentioned in Sec. 1), whereas the encoders capture a dependence relation between X and Z. Accordingly, we introduce a probability measure $Q_{X,Z}$ on

---

[1] In the appendix, we carefully construct $\mathbf{L}^2$-spaces as sets of equivalence classes of functions (see Appx. B.1) for explicitness and rigor. We do not belabor this distinction in the main text.

$\mathcal{X} \times \mathcal{Z}$, called the *pre-training distribution*, and the *prompt distribution* $\rho_{Y,Z}$ on $\mathcal{Y} \times \mathcal{Z}$ which represents the user-defined strategy for generating prompts. As a theoretical model for ZSP, we propose the function

$$\eta_\rho(\boldsymbol{x}) = \mathbb{E}_{Q_{X,Z}} \left[g_\rho(Z)|X \right] (\boldsymbol{x}), \qquad (4)$$

called the *indirect predictor*, where

$$g_\rho(\boldsymbol{z}) = \mathbb{E}_{\rho_{Y,Z}} \left[r(Y)|Z \right] (\boldsymbol{z}). \qquad (5)$$

Notice that $\eta_\star$ relies only on $P_{X,Y}$ while η_ρ is a two-stage predictor relying only on the pair $(Q_{X,Z}, \rho_{Y,Z})$. The pre-training, evaluation, and prompt distributions represent pairwise dependencies between the random variables X, Y, and Z, as well as the observable data of the problem. Intuitively, our analysis of the performance gap between the direct and indirect predictors will quantify the "compatibility" of these three fundamental distributions as a possible joint distribution on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$.

Prediction Paths of FSL and ZSP. For context, we contrast our setup with previous theoretical analyses of FSL, aiming to 1) highlight the fundamental differences between SSL-for-FSL and SSL-for-ZSP, 2) describe assumptions we make (and do not make) to best align with applications. First, we consider two common SSL tasks that precede FSL. In unimodal contrastive learning, X and Z are augmented/corrupted images, and the pretext task is to identify examples derived from the same ("+") or different ("−") underlying image (Chen et al., 2020). In reconstructive SSL, the encoder is pre-trained to predict a hidden portion of the raw/embedded image (Assran et al., 2023). Foundational works such as Saunshi et al. (2019)

and Wang and Isola (2020) explain the success of these SSL-for-FSL pipelines by the following mechanism: the labels $\mathcal{Y}$ used in the downstream task form a latent variable mixture model for the pre-training set, i.e. $Q_{X,Z} = \sum_{\boldsymbol{y} \in \mathcal{Y}} Q_{X,Z|Y=\boldsymbol{y}} \cdot Q_Y(\boldsymbol{y})$. Thus, generalization guarantees hinge upon the fact that learning parameters of the pre-training distribution must inherently capture its latent variables (the downstream labels). This theory is visualized in Fig. 1 (left & center); observe that if the dotted arrows were absent, the only path to solve the pretext task is *through* the label. This FSL "prediction path" motivates another prevalent assumption of exact/approximate conditional independence of X and Z given Y (e.g., as in Lee et al. (2021)). We avoid this assumption, which is unrealistic in the multimodal context as the dependence between an image and its caption is unlikely to be fully explained by a coarse label such as "cat". Moreover, this *latent label model* assumes equality of the marginals $P_X = Q_X$ on $\mathcal{X}$. As a concrete example, this amounts to assuming that the marginal distribution of images on the Internet (Q_X) is equal to that of CIFAR-10 images (P_X). We explicitly track this mismatch in our generalization bounds.

For ZSP, the prevailing SSL pretext task is multimodal contrastive learning (Fig. 1, right), wherein the foundation model learns a similarity function $(\boldsymbol{x}, \boldsymbol{z}) \mapsto \langle \boldsymbol{\alpha}(\boldsymbol{x}), \boldsymbol{\beta}(\boldsymbol{z}) \rangle$. To discuss a joint distribution $P \equiv P_{X,Y,Z}$, we adopt a *latent caption model* that associates $X \sim P_X$ with an unobserved $\mathcal{Z}$-valued latent variable Z (i.e. an unobserved caption). Because pre-training connects X to Z and prompting then connects Y to Z, the ideal dependence structure for ZSP is fundamentally different from FSL; if X and Y are conditionally independent given Z, the direct and indirect predictors are in fact equivalent. Indeed, the tower property of conditional expectation gives the identity

$$\begin{aligned}
\eta_\star(\boldsymbol{x}) &= \mathbb{E}_P\left[r(Y)|X\right](\boldsymbol{x}) \\
&= \mathbb{E}_P\left[\mathbb{E}_P\left[r(Y)|Z,X\right]|X\right](\boldsymbol{x}) \\
&= \mathbb{E}_P\left[\mathbb{E}_P\left[r(Y)|Z\right]|X\right](\boldsymbol{x}). \quad (X \perp\!\!\!\perp Y|Z)
\end{aligned}$$

The final expression is not equal to (4) because of the difference between $(Q_{X,Z}, \rho_{Y,Z})$ and $(P_{X,Z}, P_{Y,Z})$. Additionally, X and Y are not necessarily conditionally independent given Z. These discrepancies are precisely exposed in our analysis via a measure of distribution mismatch and a measure of the conditional dependence of X and Y given Z. The latter formalizes the information-theoretic cost of using natural language as a proxy for image classification.

Representations of the Indirect Predictor. We establish several central identities involving the indirect predictor (4). These expressions will strengthen the justification for η_ρ as the target function of ZSP and naturally lead to two classes of learning methods that we analyze in Sec. 3.

As a preview, consider the example of balanced binary classification ($r(\boldsymbol{y}) = \mathbb{1}\{\boldsymbol{y} = 1\}$) and the classifier that returns 1 when $\eta_\rho(\boldsymbol{x}) \geq 1/2$ and 0 otherwise. We will show that there exist encoders $\boldsymbol{\alpha} : \mathcal{X} \to \mathbb{R}^d$ and $\boldsymbol{\beta} : \mathcal{Z} \to \mathbb{R}^d$, and a sequence of scalars $\sigma_1 \geq \ldots \geq \sigma_d \geq 0$ such that if $\rho_Z \approx Q_Z$ and d is sufficiently large, then this classifier is equivalent to

$$\boldsymbol{x} \mapsto \arg\max_{\boldsymbol{y} \in \mathcal{Y}} \left\langle \boldsymbol{\alpha}(\boldsymbol{x}), \mathbb{E}_{\rho_{Y,Z}}\left[\boldsymbol{\beta}(Z)|Y=\boldsymbol{y}\right] \right\rangle_\sigma, \quad (6)$$

where $\langle \boldsymbol{u}, \boldsymbol{v} \rangle_\sigma := \sum_{i=1}^d \sigma_i u_i v_i$. This expression mirrors (1) down to a rescaling of the inner product. We now present the expressions that are used to derive (6).

For the first, let Q_X and Q_Z be the marginals of $Q_{X,Z}$ on $\mathcal{X}$ and $\mathcal{Z}$, respectively. We introduce the fundamental *conditional mean operator* $\mathbf{M}_{Z|X} : \mathbf{L}^2(Q_Z) \to \mathbf{L}^2(Q_X)$, which assigns to any $g \in \mathbf{L}^2(Q_Z)$ the function $\boldsymbol{x} \mapsto \mathbb{E}_{Q_{X,Z}}\left[g(Z)|X\right](\boldsymbol{x})$. Then, it holds by definition that

$$\eta_\rho(\boldsymbol{x}) = [\mathbf{M}_{Z|X} g_\rho](\boldsymbol{x}). \quad (7)$$

For the second, consider the case in which $Q_{X,Z} \ll Q_X Q_Z$[2], where $Q_X Q_Z$ denotes the probability distribution of the pair (X, Z) drawn independently as $X \sim Q_X$ and $Z \sim Q_Z$. Then, we define the Radon-Nikodym derivative $\mathsf{R} := \frac{\mathrm{d}Q_{X,Z}}{\mathrm{d}Q_X Q_Z} : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$. The function R, called the *information density*[3] has a long history in statistics and information theory. Using R (Lem. 6, Appx. B.3), the indirect predictor writes as

$$\begin{aligned}
\eta_\rho(\boldsymbol{x}) &= \mathbb{E}_{Q_Z}\left[g_\rho(Z)\mathsf{R}(\boldsymbol{x}, Z)\right] \\
&= \mathbb{E}_{\rho_{Y,Z}}\left[r(Y)\mathsf{R}(\boldsymbol{x}, Z)\right] + \mathrm{err}(Q_Z, \rho_Z), \quad (8)
\end{aligned}$$

where $\mathrm{err}(Q_Z, \rho_Z)$ term measures the discrepancy between the marginal distributions of the captions generated during pre-training and prompting, respectively. The expressions (7) and (8), while equal at the population level, motivate two categories of approaches for learning/estimation that have different statistical properties. The "conditional mean" approach uses pre-training data to learn the operator $\mathbf{M}_{Z|X}$ and prompts to approximate the function g_ρ. On the other hand, the "information density" approach learns the function R during pre-training, and approximates the expectation over $\rho_{Y,Z}$ using prompts. The information density approach is particularly reflective of the prompting aspect of (1), as one may perceive $\boldsymbol{z}_k^{\boldsymbol{y}}$ for $k = 1, \ldots, m$ and $\boldsymbol{y} \in \mathcal{Y}$ as $M = m|\mathcal{Y}|$ as samples from

---

[2]A distribution μ on $\mathcal{U}$ is *absolute continuous* with respect to another distribution ν (denoted $\mu \ll \nu$) if $\nu(A) = 0 \implies \mu(A) = 0$ for every measurable set $A \subseteq \mathcal{U}$. If so, there exists a *Radon-Nikodym derivative* $\frac{\mathrm{d}\mu}{\mathrm{d}\nu} : \mathcal{U} \to \mathbb{R}_{\geq 0}$ such that for every measurable set $A \subseteq \mathcal{U}$, it holds that $\mu(A) = \int_A \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\boldsymbol{u})\mathrm{d}\nu(\boldsymbol{u})$.

[3]This term actually refers to $(\boldsymbol{x}, \boldsymbol{z}) \mapsto \log \mathsf{R}(\boldsymbol{x}, \boldsymbol{z})$, but for simplicity, we use it for R—see Dytso et al. (2023), Eq. (11)).

$\rho_{Y,Z}$ with ρ_Y chosen to be uniform on $\mathcal{Y}$. These are then used to replace the expectation in (8).

Finally, we tie back to (6) and describe the formal connection between $\mathbf{M}_{Z|X}$ and R. In Prop. 2 (Appx. B.3), we prove the decomposition of the form

$$R(\boldsymbol{x}, \boldsymbol{z}) = \langle \boldsymbol{\alpha}(\boldsymbol{x}), \boldsymbol{\beta}(\boldsymbol{z}) \rangle_\sigma + \varepsilon_d, \qquad (9)$$

where $\sigma_d, \varepsilon_d \to 0$ as $d \to \infty$. Then, (6) follows under the given conditions by plugging (9) into (8).

The encoders $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$, and constants $(\sigma_i)_{i=1}^d$ are none other than the components of the truncated singular value decomposition (SVD) of $\mathbf{M}_{Z|X}$ (Prop. 1, Appx. B.3). The SVD of $\mathbf{M}_{Z|X}$ and the information density R characterize the full dependence structure of $Q_{X,Z}$; because R is identically 1 when $Q_{X,Z} = Q_X Q_Z$, we may define the (squared) *mean square contingency* dependence measure

$$I(X; Z) = \mathbb{E}_{Q_X Q_Z} \left[(R(X, Z) - 1)^2 \right] \qquad (10)$$
$$= \|\mathbf{M}_{Z|X}\|_{\mathrm{HS}}^2 - 1$$
$$= \sum_{i=2}^{\infty} \sigma_i^2, \qquad (11)$$

where $\|\cdot\|_{\mathrm{HS}}$ denotes the Hilbert-Schmidt norm (see Definition 7, Appx. B.2), and the identities are proven in Prop. 2. The right-hand side of (10) can also be interpreted as the χ^2-divergence $D_{\chi^2}(Q_{X,Z} \| Q_X Q_Z) := \mathbb{E}_{Q_X Q_Z}[(\frac{\mathrm{d}Q_{X,Z}}{\mathrm{d}Q_X Q_Z}(X, Z) - 1)^2]$ between the joint distribution and the product of the marginals (see Definition 8).

3. Generalization Guarantees for ZSP

In this section, we prove generalization guarantees for ZSP methods by comparing $\eta_\star$ to η_ρ and η_ρ to an estimator $\hat{\eta}_\rho$, based on an N-sized pre-training set and M-sized prompt set (recall that $M = m|\mathcal{Y}|$ in (1)). While there are some subtleties in the sampling models between various methods, one can consider $(X_1, Z_1), \ldots, (X_N, Z_N) \overset{\text{i.i.d}}{\sim} Q_{X,Z}$ and $(Y_1, Z_1'), \ldots, (Y_M, Z_M') \overset{\text{i.i.d}}{\sim} \rho_{Y,Z}$ for intuition purposes (see Appx. D.5 for a detailed description). We consider specific instances of both the conditional mean and information density approaches, based on learning theory in reproducing kernel Hilbert space (RKHS); our arguments do not intend to interpret foundation modeling as a kernel method, but to use the detailed analysis of the statistical errors in kernel methods to gain insight. In particular, we aim to expose two key dependences for the random triple (X, Y, Z): the dependence between X and Z (which governs pre-training) and the conditional dependence between X and Y given Z (which governs downstream prediction). Similar statistical guarantees for other function classes (reviewed in Appx. E) can be plugged into our framework,

which intends to capture the end-to-end performance from pre-training to downstream prediction.

For $h \in \mathbf{L}^2(P_X)$, we define the norm $\|h\|_{\mathbf{L}^2(P_X)}^2 := \int_{\mathcal{X}} h^2(\boldsymbol{x}) \mathrm{d}P_X(\boldsymbol{x})$, using analogous notation for other probability distributions. We will assume throughout the paper that r is bounded by B_r with probability one under P_Y and ρ_Y, so that $\eta_\rho, \eta_\star \in \mathbf{L}^2(P_X)$. Given a square-integrable $\hat{\eta}_\rho$, we first control the mean squared error (MSE) via $\|\eta_\star - \hat{\eta}_\rho\|_{\mathbf{L}^2(P_X)}^2 \leq$

$$2 \underbrace{\|\eta_\star - \eta_\rho\|_{\mathbf{L}^2(P_X)}^2}_{\text{information-theoretic error}} + 2 \underbrace{\|\eta_\rho - \hat{\eta}_\rho\|_{\mathbf{L}^2(P_X)}^2}_{\text{estimation error}}. \qquad (12)$$

The information-theoretic error captures the prompt bias and residual dependence that differentiates indirect and direct prediction, whereas the estimation error is a familiar term in statistical analysis. We discuss in Appx. D.4 how to convert the MSE bounds to risk bounds for classification.

Prompt Bias and Residual Dependence. Here, we control the information-theoretic error term in (12). We state our assumptions regarding conditional probability informally and defer the formal descriptions using the language of regular conditional distributions to Appx. C. We work within the latent caption model from Sec. 2, for which we consider a joint distribution $P_{X,Y,Z}$ on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ which equals the evaluation distribution $P_{X,Y}$ when marginalized over $\mathcal{Z}$. Similar to the information density R from Sec. 2, we introduce the conditional information density

$$S_{\boldsymbol{z}} := \frac{\mathrm{d}P_{X,Y|\boldsymbol{z}}}{\mathrm{d}(P_{X|\boldsymbol{z}} P_{Y|\boldsymbol{z}})} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}, \qquad (13)$$

where $P_{X,Y|\boldsymbol{z}}$ denotes the conditional distribution of (X, Y) given $Z = \boldsymbol{z}$, and $P_{X|\boldsymbol{z}} P_{Y|\boldsymbol{z}}$ is defined analogously. This naturally motivates the conditional dependence measure given by

$$I(X; Y|\boldsymbol{z}) = \mathbb{E}_{P_{X|\boldsymbol{z}} P_{Y|\boldsymbol{z}}} \left[(S_{\boldsymbol{z}}(X, Y) - 1)^2 \right], \qquad (14)$$

called the *conditional mean square contingency*. Finally, consider the following regularity assumption on the joint distribution $P_{X,Y,Z}$, also discussed in Appx. C.

Assumption 1. $P_{X,Y,Z}$ on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ satisfies the following: *1) Agreement of caption distribution:* P_X-almost all $\boldsymbol{x} \in \mathcal{X}$, $P_{Z|\boldsymbol{x}}$ exists and $P_{Z|\boldsymbol{x}} = Q_{Z|\boldsymbol{x}}$. *2) Regularity of conditional distributions:* For P_Z-almost all $\boldsymbol{z} \in \mathcal{Z}$, $P_{X,Y|\boldsymbol{z}}$ exists, $P_{X,Y|\boldsymbol{z}} \ll P_{X|\boldsymbol{z}} P_{Y|\boldsymbol{z}}$, and the conditional information density (13) satisfies $\mathbb{E}_{P_{X,Y|\boldsymbol{z}}}[S_{\boldsymbol{z}}(X, Y)] < +\infty$ and $\mathbb{E}_{P_{X,Y,Z}}[S_Z(X, Y)] < +\infty$.

To measure the bias of the prompt distribution $\rho_{Y,Z}$, we

denote the analog of (5) under $P_{Y,Z}$ as

$$g_{P_{Y,Z}}(\boldsymbol{z}) = \mathbb{E}_{P_{Y,Z}}\left[r(Y)|Z\right](\boldsymbol{z}).$$

We may now state the main result, proved in Appx. C.

Theorem 1. *Under Asm. 1,* $\|\eta_\rho - \eta_\star\|^2_{\mathbf{L}^2(P_X)} \lesssim$

$$\underbrace{\mathbb{E}_{P_Z}\left[I(X;Y|Z)\right]}_{\text{residual dependence}} + \underbrace{\|g_\rho - g_{P_{Y,Z}}\|^2_{\mathbf{L}^2(P_Z)}}_{\text{prompt bias}}. \quad (15)$$

To give context to Thm. 1, conditional independence relations have previously been used to describe the performance of multimodal contrastive SSL for FSL. We are particularly inspired by the *multi-view redundancy* theory of Tosh et al. (2021), which states informally that the population FSL predictor can approach the performance of the idealized direct predictor that is given *both* X and Z at test time, if $X \perp\!\!\!\perp Y|Z$ and $Z \perp\!\!\!\perp Y|X$ approximately hold. However, the theory of graphical models (Lauritzen, 1996, Proposition 3.1) asserts that both conditional independence relations hold only if $(X, Z) \perp\!\!\!\perp Y$, that is, neither view has information predictive of the label. This can be seen intuitively from Fig. 1 by breaking the arrows $X \rightarrow Y$ and $Z \rightarrow Y$. Notice that we compare only to the direct predictor (3) given X (which is reflective of practice), so that we need only that $X \perp\!\!\!\perp Y|Z$ (i.e. X depends on Y through Z) to close the gap. The prompt bias term (15) captures the possible incompatibility of the prompt distribution $\rho_{Y,Z}$ with $(P_{X,Y}, Q_{X,Z})$—we call prompt strategies unbiased (see Appx. D.5) when this term is zero.

Sample Complexity and Distribution Mismatch. The first step in our estimation error analysis is to pass the $\mathbf{L}^2(P_X)$-norm term $\|\eta_\rho - \hat{\eta}_\rho\|^2_{\mathbf{L}^2(P_X)}$ from (12) to the $\mathbf{L}^2(Q_X)$-norm counterpart $\|\eta_\rho - \hat{\eta}_\rho\|^2_{\mathbf{L}^2(Q_X)}$. We then establish high-probability bounds on the $\mathbf{L}^2(Q_X)$-norm term, with respect to the random sampling of the pre-training and prompting data. Because this initial step follows from a standard distribution shift argument (based on either a bounded likelihood ratio assumption or an additive error in total variation distance), we defer it to Appx. D (see Lem. 14). Conceptually, the two examples below are derived from estimating the component of either (7) or (8) that involves $Q_{X,Z}$ using the pre-training set and the one that involves $\rho_{Y,Z}$ using the prompt strategy. In both cases, we discuss the convergence rates of state-of-the-art RKHS-based methods. As we review Appx. B.4, these rates are typically expressed in terms of two quantities: *source condition* constants, which measure the smoothness of the target function being learned, and *eigendecay exponents* of covariance operators, which measure the effective dimension of the data. It will serve our purposes to interpret the rates in terms of the dependence between X and Z under

$Q_{X,Z}$, under the following assumption.

Assumption 2. The pre-training distribution satisfies $Q_{X,Z} \ll Q_X Q_Z$, and the information density R is contained in $\mathbf{L}^2(Q_X Q_Z)$ (i.e. $I(X;Z)$ is well-defined).

Due to the technical overhead of each method (especially regarding mis-specified function classes), we provide high-level descriptions below and defer detailed descriptions of the specific estimation procedures and formal assumptions to Appx. D.1 (conditional mean) and Appx. D.2 (information density). We denote by $\delta \in (0, 1]$ a failure probability, and $\text{plog}(\cdot)$ a term that is poly-logarithmic in its input.

Example 1: Nonparametric Regression. This approach, based on (7), uses the pre-training set to produce an estimate $\widehat{\mathbf{M}}_{Z|X}$ of the conditional mean operator and the prompts to produce an approximation $\hat{g}_\rho : \mathcal{Z} \rightarrow \mathbb{R}$ of g_ρ. For the former, we use as an example the spectral regularization learning method of Meunier et al. (2024), which produces a conditional mean embedding function $\widehat{F} : \mathcal{X} \rightarrow \mathcal{G}$, for an RKHS $\mathcal{G}$ of real-valued functions of $\mathcal{Z}$. For any $g \in \mathcal{G}$, we then define $[\widehat{\mathbf{M}}_{Z|X}g](\boldsymbol{x}) = \langle g, \widehat{F}(\boldsymbol{x})\rangle_{\mathcal{G}}$. Note that $\widehat{F}$ predicts a target that is itself a function–such methods are therefore referred to as "vector-valued" regression. By the Reisz representation theorem, a similar function $F_\star$ can be constructed such that $[\mathbf{M}_{Z|X}g](\boldsymbol{x}) = \langle g, F_\star(\boldsymbol{x})\rangle_{\mathcal{G}}$. For $\hat{g}_\rho$, we consider standard kernel regularized least-squares (e.g., Smale and Zhou (2007)) applied to M i.i.d. draws from $\rho_{Y,Z}$. Assuming that $g_\rho \in \mathcal{G}$, one can then pass the problem to controlling $\|\hat{g}_\rho - g_\rho\|^2_{\mathcal{G}}$ and $\|\widehat{F} - F_\star\|^2_{\mathbf{L}^2(Q_X;\mathcal{G})}$, where $\mathbf{L}^2(Q_X;\mathcal{G})$ denotes a Bochner space (reviewed in Appx. B.4).

To derive the convergence rates below, we show in Appx. D.1 that the source condition on $F_\star$ can be expressed in terms of the singular decay exponent of $\mathbf{M}_{Z|X}$ (i.e. $\sigma_i \sim i^{-\gamma_{X,Z}}$ from (11)), and the eigendecay exponents γ_X and γ_Z of the covariance operators of Q_X and Q_Z, respectively. Additionally, $\omega_\rho > 1/2$ is a parameter controlling the convergence rate of the prompt-based estimate of g_ρ. The parametrization below is chosen so that one may interpret ω_ρ as a similar source condition for the target function g_ρ. In the well-specified case (when $F_\star$ is contained in the hypothesis class), we describe the convergence rate with the aggregated exponent

$$q(t) = (2\gamma_{X,Z} + \gamma_Z - 1)^t \gamma_X^{1-t} \geq 1, \quad t \in [0, 1)$$

where t depends on $F_\star$. The result below corresponds to Thm. 10 in Appx. D.1, which relies on a basis alignment assumption to aggregate the singular/eigendecays.

Theorem 2 (Informal). *For* $\hat{\eta}_\rho(\boldsymbol{x}) = \langle \hat{g}_\rho, \widehat{F}(\boldsymbol{x})\rangle_{\mathcal{G}}$, *there exist* $t \in [0, 1)$ *and* $C(Q_{X,Z}) \geq 0$ *(independent of*

(N, M, δ)) such that $\|\hat{\eta}_\rho - \eta_\rho\|^2_{\mathbf{L}^2(Q_X)} \lesssim$

$$\text{plog}(1/\delta)\left[N^{-\frac{q(t)}{q(t)+1}} + C(Q_{X,Z})M^{-\frac{2\omega_\rho-1}{2\omega_\rho+1}}\right] \qquad (16)$$

with probability at least $1 - \delta$ for N sufficiently large.

Let us interpret the constant $q(t)$. First, the dependence on N ranges between $O(N^{-1/2})$ and the parametric rate $O(N)$. Convergence is faster when $\gamma_{X,Z} \gg 1$ or $\gamma_Z \gg 1$. The first case implies near-independence of X and Z, for which learning is easy as $\widehat{F}(\boldsymbol{x})$ is essentially constant over $\boldsymbol{x} \in \mathcal{X}$. The second case indicates that the Z variable is near-finite dimensional, or that the vector-valued nature of the problem has been reduced to standard univariate regression. Convergence is slower if $\gamma_X \gg 1$, or if the effective dimension of $\mathcal{X}$ is small relative to the effective dimension of Z. The balancing constant $C(Q_{X,Z})$ (shown explicitly in Thm. 10) decays with $\gamma_{X,Z}$ and γ_Z, so as (X, Z) becomes more independent or Z approaches finite dimensions, the variance from prompt sampling decreases. We also discuss the mis-specified case in Appx. D.1.

Example 2: Radon-Nikodym Derivative Estimation.
This approach, based on (8), considers pre-training to return a learned information density $\widehat{\mathsf{R}} : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$. By approximating the prompt distribution $\rho_{Y,Z}$ with $\hat{\rho}_{Y,Z}$ (e.g. the empirical measure in the result below), one may define the estimator $\hat{\eta}_\rho(\boldsymbol{x}) = \mathbb{E}_{\hat{\rho}_{Y,Z}}[r(Y)\widehat{\mathsf{R}}(\boldsymbol{x}, Z)]$. Similar in spirit to the previous example, we consider the kernel Radon-Nikodym derivative estimation with the spectral regularization procedure of Nguyen et al. (2024). The convergence rate of $\widehat{\mathsf{R}}$ to R is governed by a source condition constant $\beta \geq 1$ associated to R (see Appx. D.2). We interpret this constant analogously to $q(t)$, in that we prove a relationship to the singular decay exponent $\gamma_{X,Z}$, but is not directly expressible in terms of the latter. The following result corresponds to Thm. 11 in Appx. D.2.

Theorem 3 (Informal)**.** *For $\hat{\eta}_\rho(\boldsymbol{x}) = \mathbb{E}_{\hat{\rho}_{Y,Z}}[r(Y)\widehat{\mathsf{R}}(\boldsymbol{x}, Z)]$, and $\rho_Z \ll Q_Z$, there exists $C_{\mathsf{R},\rho}(Q_X) \geq 0$ (independent of (N, M, δ)) such that $\|\hat{\eta}_\rho - \eta_\rho\|^2_{\mathbf{L}^2(Q_X)} \lesssim$*

$$\text{plog}(1/\delta)\left[N^{-\frac{\beta}{\beta+1}} + C_{\mathsf{R},\rho}(Q_X)M^{-1}\right] + D_{\chi^2}(\rho_Z\|Q_Z)$$

with probability at least $1 - \delta$ for all N sufficiently large.

Notice that the bound of Thm. 3 includes a divergence term between ρ_Z (the captions generated by prompting) and Q_Z (the captions of the pre-training set). This term comes precisely from the error term in (8). This elucidates the fact that the conditional mean approach and the information density are not equivalent representations of the pre-training distribution, as one needs *both* R and Q_Z in order to identify the conditional mean. The parametric

rate M^{-1} reflects that samples are used to learn a joint expectation over $\rho_{Y,Z}$, which is an easier statistical problem than estimating the regression function of Y on Z that appears in Thm. 2. Thus, the information density approach may enjoy faster statistical convergence, at the expense of bias from the distribution mismatch on $\mathcal{Z}$. The constant $C_{\mathsf{R},\rho}(Q_X)$ relates to the $\mathbf{L}^2(Q_X)$-norm of the random function $\boldsymbol{x} \mapsto r(Y)\mathsf{R}(\boldsymbol{x}, Z)$ for $(Y, Z) \sim \rho_{Y,Z}$; the error from finite prompts decays when this norm is light-tailed.

In both Thm. 2 and Thm. 3, we aim to highlight not particular convergence rates of the chosen methods, but the framework that leads to proving them. Similar results can also be leveraged in our framework. SSL procedures such as noise contrastive estimation have been related to the estimation of R (Gutmann and Hyvärinen, 2012). For example, Tosh et al. (2021, Theorem 11) upper bounds $\|\widehat{\mathsf{R}} - \mathsf{R}\|^2_{\mathbf{L}^2(Q_XQ_Z)}$ using the suboptimality of the population risk, allowing for empirical risk minimization-style analysis.

4. Experiments

In Sec. 1, we asked how the downstream task performance depends on the pre-training distribution $Q_{X,Z}$, evaluation distribution $P_{X,Y}$, and prompting strategy $\rho_{Y,Z}$. At the population level, we captured the dependence on $Q_{X,Z}$ and $P_{X,Y}$ using the residual dependence $\mathbb{E}_{P_Z}[I(X; Z)]$ and incorporated $\rho_{Y,Z}$ via the prompt bias (Thm. 1). In the first experiment, we create a simulated setting in which the residual dependence can be controlled and investigate whether it is indeed a determining factor for the empirical performance of CLIP (Radford et al., 2021) and VICReg (Bardes et al., 2022) models in practice. In the second experiment, we solve an image classification task in which the images have both captions and labels (i.e., we may sample from a true joint distribution $P_{X,Y,Z}$). This allows us to understand the effect of prompt bias by comparing template-based prompting strategies to the unbiased setting $\rho_{Y,Z} = P_{Y,Z}$. To understand the dependence on $\rho_{Y,Z}$ at a sample level, we explore how downstream performance scales with the number of prompts M in both the second experiment (unbiased prompting) and third experiment (LLM-based prompting). We are particularly interested in verifying the dependence on M (which is the dominant error when $N \gg M$) derived in Thm. 3. Appx. F contains further details of the experiments, and code for reproduction can be found at `github.com/ronakdm/zeroshot`.

Models, Datasets, and Evaluation. For foundation models, we use three publicly available CLIP models from the OpenCLIP repository (Ilharco et al., 2022): ResNet50 pre-trained on YFCC15M (Thomee et al., 2016), NLLB-CLIP pre-trained on a subset of LAION COCO (Visheratin, 2023), and ViT-B/32 pre-trained on the DataComp
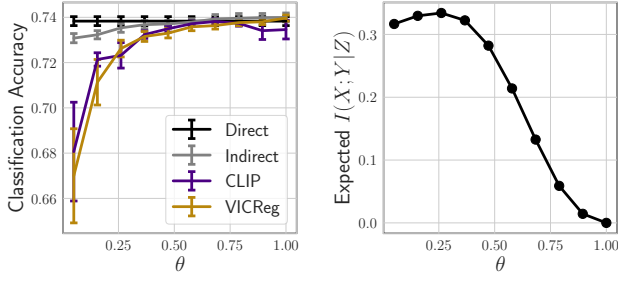
Figure 2. **Results: Residual Dependence Simulation.** Simulation for (X, Z, Y) described in Appx. F.4. **Left:** The y-axis is the accuracy of classifying Y given X and the x-axis is the parameter θ controlling the residual dependence $I(X;Y|Z)$ as in (103). **Right:** The y-axis shows $\mathbb{E}_{P_Z}[I(X;Y|Z)]$ as computed in Appx. F.4. Error bars indicate standard errors from 10 seeds, which govern the data used for estimating expected values and randomness in the training procedures for CLIP and VICReg.

medium pool (Gadre et al., 2023). Our evaluation datasets include five standard benchmarks: the Describable Textures Dataset or DTD (Cimpoi et al., 2014), Flowers 102 (Nilsback and Zisserman, 2008), FGVC Aircraft (Maji et al., 2013), SUN397 (Xiao et al., 2010), and ImageNet-1k (Deng et al., 2009). For some experiments, we make use of the ImageNet-Captions dataset (Fang et al., 2023), which pairs a subset of ImageNet images collected from Flickr with their original captions. Evaluation occurs via zero-shot classification top-k accuracy, in which a test example is considered to be classified correctly if the true class is contained within the elements of $\mathcal{Y}$ with the k largest scores as computed by (1). Evaluation is done using tools from the CLIP Benchmark repository. In Fig. 3 and Fig. 4, "templates" refers to using all default community-curated prompts available in CLIP Benchmark. Finally, detailed descriptions of the prompt sampling schemes are collected and compared to the theory in Appx. D.5.

Classification Accuracy and Residual Dependence. We consider a simulated binary classification task in which all distributions are compatible (i.e. $Q_{X,Z} = P_{X,Z}$ and $\rho_{Y,Z} = P_{Y,Z}$ for some $P_{X,Y,Z}$) and the predictors (3) and (4) can be computed analytically. We also include the zero-shot predictor (1) learned by both the CLIP and VICReg objectives. Our goals are two-fold in this simulation: 1) to empirically observe that as $\mathbb{E}_{P_Z}[I(X;Y|Z)] \to 0$, the predictive performance of the indirect predictor η_ρ does indeed approach that of $\eta_\star$, and 2) that the predictors generated by common SSL methods used in practice have similar performance trends as η_ρ. As for the data-generating process, we consider $\mathcal{X} = \mathcal{Z} = \mathbb{R}^d$ and a pair of Gaussian
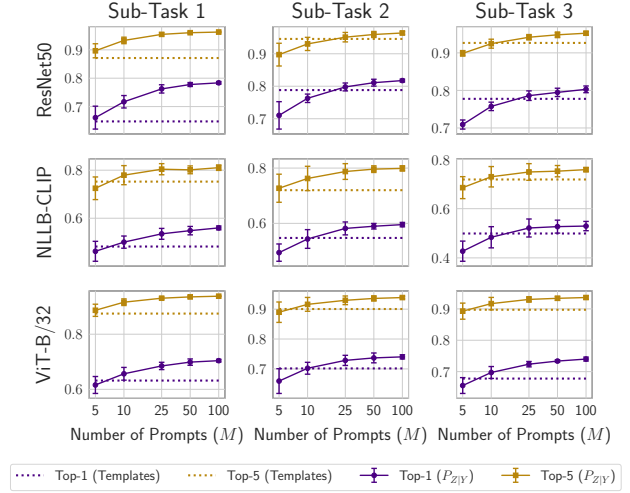

Figure 3. **Results: Unbiased Prompting.** Pre-trained models are varied along the rows and sub-tasks (subsets of 50 ImageNet-1k classes) are varied along columns. In all plots, the x-axis denotes the number of prompts sampled for each class embedding and the y-axis denotes top-k zero-shot classification accuracy. Error bars indicate standard deviations across 10 seeds for prompt sampling.

distributions $(P_{X,Z|Y=0}, P_{X,Z|Y=1})$, where given $Y = \boldsymbol{y}$,

$$\begin{bmatrix} X \\ Z \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_{X|\boldsymbol{y}} \\ \boldsymbol{\mu}_{Z|\boldsymbol{y}} \end{bmatrix}, \begin{bmatrix} \mathbf{C}_{XX|\boldsymbol{y}} & \mathbf{C}_{XZ|\boldsymbol{y}} \\ \mathbf{C}_{ZX|\boldsymbol{y}} & \mathbf{C}_{ZZ|\boldsymbol{y}} \end{bmatrix} \right)$$

with class-conditional mean vectors $\boldsymbol{\mu}_{X|\boldsymbol{y}}, \boldsymbol{\mu}_{Z|\boldsymbol{y}} \in \mathbb{R}^d$ and covariance matrices $\mathbf{C}_{XX|\boldsymbol{y}}, \mathbf{C}_{ZX|\boldsymbol{y}}, \mathbf{C}_{ZZ|\boldsymbol{y}} \in \mathbb{R}^{d \times d}$. In order to control the conditional dependence between X and Y given Z, we fix all parameters except for $\boldsymbol{\mu}_{Z|\boldsymbol{y}}$ and $\mathbf{C}_{ZX|\boldsymbol{y}}$ (for $\boldsymbol{y} = 0, 1$), and define them using a tunable parameter $\theta \in [0, 1]$ in a way such that the conditional distribution of Y given $X = \boldsymbol{x}$ stays constant. We make it so that as $\theta \to 1$, $I(X;Y|\boldsymbol{z}) \to 0$. Finally, to measure classification accuracy, we directly draw samples from $P_{Y,Z}$ to simulate unbiased prompting. The full mathematical details are given in Appx. F.4. We observe both of the intended effects; the left panel of Fig. 2 demonstrates that as θ approaches 1, the indirect, CLIP, and VICReg predictors approach the performance of the direct predictor in terms of classification performance. The right panel confirms that θ indeed controls $\mathbb{E}_{P_Z}[I(X;Y|Z)]$ in an approximately monotonic fashion.

Prompting without Bias with Observations from $P_{Z,Y}$. Next, we illustrate the importance of the prompt bias term in Thm. 1 by considering an ImageNet-Captions dataset, in which we may observe the joint sample (X, Y, Z). We compare the standard prompting technique using predefined templates to the unbiased strategy that draws samples directly from $P_{Y,Z}$. We design three sub-tasks by
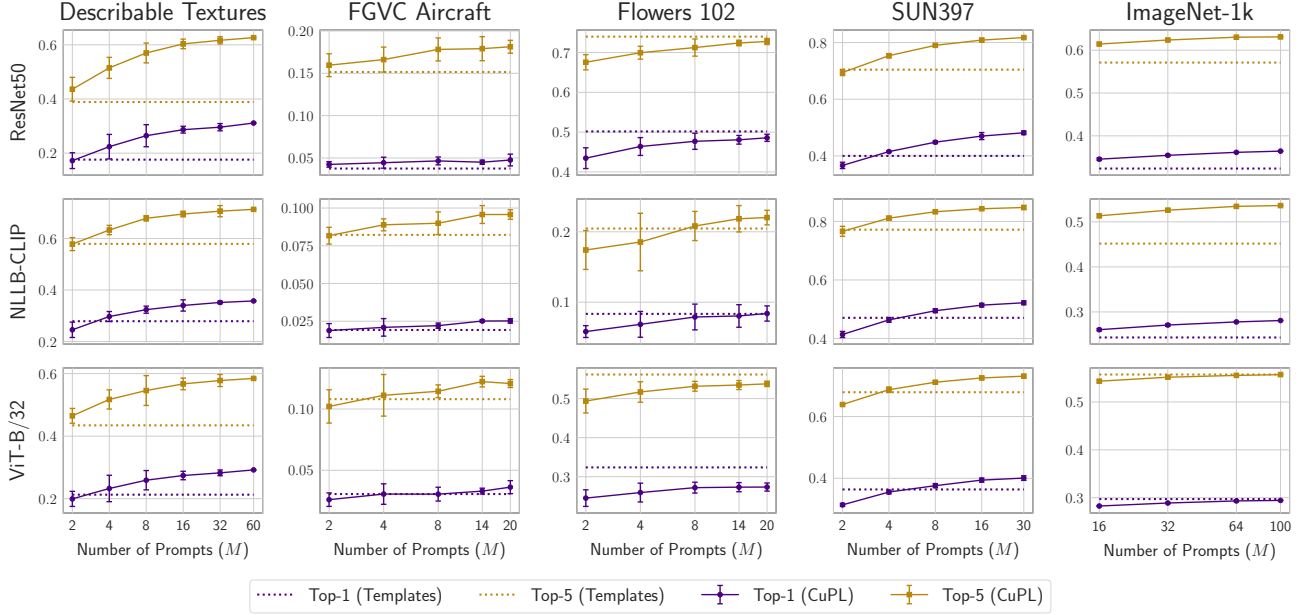
Figure 4. **Results: Class-Conditional Prompting.** Pre-trained models are varied along the rows and evaluation datasets are varied along columns. In all plots, the x-axis denotes the number of prompts sampled for each class embedding and the y-axis denotes top-k zero-shot classification accuracy. Error bars indicate standard deviations across 10 seeds for prompt sampling.

randomly selecting collections of 50 classes from each of 998 classes, reserving held-out prompting examples for which we can draw from $P_{Z|Y=y}$ for each $y \in \mathcal{Y}$ (see the additional details in Appx. F). The zero-shot classification accuracy on a held-out evaluation set is plotted in Fig. 3. Observe that the threshold at which unbiased prompting outperforms the 18 default templates is approximately $M = 10$ across tasks. However, the performance of the unbiased approach only saturates at $M = 100$ and can have enormous benefits (almost 15% absolute increase in top-1 accuracy for the ResNet50 on Sub-Task 1) in performance. Thus, for models that have not yet been saturated from pre-training, prompting can close surprisingly wide gaps in zero-shot classification accuracy.

Class-Conditional Prompting with Language Models.
As mentioned in Sec. 1, we investigate CuPL as a means to implement class-conditional prompting (sampling from $\rho_{Z|Y=y}$ for each $y \in \mathcal{Y}$) with LLMs. Our experimental setup and scientific goals differ from those used in Pratt et al. (2023): 1) we use lightweight encoders that have not saturated their performance during pre-training, as opposed to the large-scale ViT-L/14 architecture, 2) we quantify the variability of classification accuracy with respect to prompting by generating up to fifty times as many prompts per experiment, and 3) we employ LlaMA 3 (Llama Team, Meta AI, 2024), which is free and accessible to other, as opposed to GPT-3 (Brown et al., 2020). The results are shown in Fig. 4, where we order the datasets in increasing

number of classes per task: 47, 100, 102, 397, and 998. Similar phenomena as in Fig. 3 are observed, although the approximate saturation threshold varies per dataset from 20 for Flowers 102 and FGVC Aircraft up to 60 for DTD. Note that the choice of defaults heavily influences the baseline performance. Surprisingly, the Flowers 102 dataset uses a single default: "a photo of a ___, a type of flower", and is often able to outperform the class-conditional LLM approach on average. On the other hand, the DTD templates of the form "a photo of a ___ {texture, pattern, thing, object}" are dramatically outperformed by our LLM-generated captions , with a nearly 20% increase in top-5 accuracy on the ResNet50 and ViT-B/32 architectures.

5. Conclusion

We showed how zero-shot prediction (ZSP) can be theoretically understood as an indirect prediction path from another modality to the label. We presented two viewpoints on categorizing ZSP methods—the conditional mean approach and the information density approach—and framed a decomposition formula for their generalization abilities. Our theoretical results and experiments highlighted the role of residual dependence and prompt bias in defining the fundamental limits of ZSP. Interesting venues for future work include the extension of our analysis to classes of distribution shifts between the pre-training distribution and the downstream distribution, and to causal generative modeling (Scetbon et al., 2024; Zhang et al., 2024).

Impact Statement

As the capabilities of foundation models become increasingly universal, we feel that a comprehensive understanding of their underlying mechanisms is crucial. Complementing the community's empirical work on the evaluating the impact of pre-training data and mitigating bias, we aim to spark an equally rigorous line of research that rethinks theoretical analysis for the purpose of modern learning paradigms that greatly affect humanity.

Acknowledgements

References

Z. Akata, Z. Harchaoui, and C. Schmid. Label-Embedding for Image Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.

M. Andrychowicz, M. Denil, S. Gómez, M. W. Hoffman, D. Pfau, T. Schaul, B. Shillingford, and N. de Freitas. Learning to learn by gradient descent by gradient descent. In *NeurIPS*, 2016.

M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. In *CVPR*, 2023.

Y. Atzmon, F. Kreuk, U. Shalit, and G. Chechik. A causal view of compositional zero-shot recognition. In *NeurIPS*, 2020.

J.-P. Aubin. *Applied Functional Analysis*. Wiley, 2nd edition, 2000.

F. Bach. *Learning Theory from First Principles*. The MIT Press, 2024.

C. R. Baker. Joint Measures and Cross-Covariance Operators. *Transactions of the American Mathematical Society*, 1973.

R. Balestriero and Y. LeCun. Contrastive and Non-Contrastive Self-Supervised Learning Recover Global and Local Spectral Embedding Methods. In *NeurIPS*, 2022.

R. Balestriero, M. Ibrahim, V. Sobal, A. Morcos, S. Shekhar, T. Goldstein, F. Bordes, A. Bardes, G. Mialon, Y. Tian, et al. A cookbook of self-supervised learning. arXiv Technical Report, 2023.

A. Bardes, J. Ponce, and Y. LeCun. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. In *ICLR*, 2022.

K. Barnard, P. Duygulu, D. Forsyth, N. d. Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *JMLR*, 2003.

F. Bartolucci, E. De Vito, L. Rosasco, and S. Vigogna. Understanding neural networks with reproducing kernel Banach spaces. *Applied and Computational Harmonic Analysis*, 2023.

F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 2007.

P. J. Bickel, C. A. Klaassen, P. J. Bickel, Y. Ritov, J. Klaassen, J. A. Wellner, and Y. Ritov. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press Baltimore, 1993.

R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. arXiv Technical Report, 2022.

L. Breiman and J. H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391): 580–598, 1985.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language Models are Few-Shot Learners. In *NeurIPS*, 2020.

A. Buja. Remarks on Functional Canonical Variates, Alternating Least Squares Methods and ACE. *The Annals of Statistics*, 1990.

V. A. Cabannnes, F. Bach, and A. Rudi. Fast Rates for Structured Prediction. In *COLT*, 2021.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

Z. Chen, Y. Deng, Y. Li, and Q. Gu. Understanding Transferable Representation Learning and Zero-shot Transfer in CLIP. In *ICLR*, 2024.

A. Christmann and I. Steinwart. *Support vector machines*. Springer, 2008.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing Textures in the Wild. In *CVPR*, 2014.

F. Cucker and D. X. Zhou. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge UP, 2007.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

R. DeVore, R. D. Nowak, R. Parhi, and J. W. Siegel. Weighted variation spaces and approximation by shallow ReLU networks. *Applied and Computational Harmonic Analysis*, 2025.

K. Du and Y. Xiang. Low-Rank Approximation of Structural Redundancy for Self-Supervised Learning. In *CLeaR*, 2024.

R. Durrett. *Probability: Theory and Examples*. Cambridge University Press, 2019.

A. Dytso, M. Cardone, and I. Zieder. Meta Derivative Identity for the Conditional Expectation. *IEEE Transactions on Information Theory*, 2023.

R. Esfandiarpoor, C. Menghini, and S. H. Bach. If CLIP Could Talk: Understanding Vision-Language Model Representations Through Their Preferred Concept Descriptions. In *EMNLP*, 2024.

A. Fang, G. Ilharco, M. Wortsman, Y. Wan, V. Shankar, A. Dave, and L. Schmidt. Data determines distributional robustness in contrastive language-image pre-training (CLIP). In *ICML*, 2023.

C. Finn, P. Abbeel, and S. Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *ICML*, 2017.

S. Fischer and I. Steinwart. Sobolev Norm Learning Rates for Regularized Least-Squares Algorithms. *JMLR*, 2020.

D. A. Forsyth, T. Berg, C. O. Alm, A. Farhadi, J. Hockenmaier, N. Loeff, and G. Wang. Words and pictures: Categories, modifiers, depiction, and iconography. *Object categorization: Computer and human vision perspectives*, 2009.

K. Fukumizu, A. Gretton, and F. Bach. Statistical Convergence of Kernel CCA. In *NeurIPS*, 2005.

K. Fukumizu, F. R. Bach, and A. Gretton. Statistical Consistency of Kernel Canonical Correlation Analysis. *JMLR*, 2007a.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel Measures of Conditional Dependence. In *NeurIPS*, 2007b.

S. Y. Gadre, G. Ilharco, A. Fang, J. Hayase, G. Smyrnis, T. Nguyen, R. Marten, M. Wortsman, D. Ghosh, J. Zhang, E. Orgad, R. Entezari, G. Daras, S. M. Pratt, V. Ramanujan, Y. Bitton, K. Marathe, S. Mussmann, R. Vencu, M. Cherti, R. Krishna, P. W. Koh, O. Saukh, A. Ratner, S. Song, H. Hajishirzi, A. Farhadi, R. Beaumont, S. Oh, A. Dimakis, J. Jitsev, Y. Carmon, V. Shankar, and L. Schmidt. DataComp: In search of the next generation of multimodal datasets. In *NeurIPS*, 2023.

I. Gohberg, S. Goldberg, and M. Kaashoek. *Classes of Linear Operators Vol. 1*. Springer, 1990.

I. Gohberg, S. Goldberg, and M. Kaashoek. *Basic Classes of Linear Operators Vol. 1*. Springer, 2003.

S. Goyal, A. Kumar, S. Garg, Z. Kolter, and A. Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *CVPR*, 2023.

Q. Guo, R. Wang, J. Guo, B. Li, K. Song, X. Tan, G. Liu, J. Bian, and Y. Yang. Connecting Large Language Models with Evolutionary Algorithms Yields Powerful Prompt Optimizers. In *ICLR*, 2024.

M. U. Gutmann and A. Hyvärinen. Noise-Contrastive Estimation of Unnormalized Statistical Models, with Applications to Natural Image Statistics. *JMLR*, 2012.

J. Z. HaoChen, C. Wei, A. Gaidon, and T. Ma. Provable Guarantees for Self-Supervised Deep Learning with Spectral Contrastive Loss. In *NeurIPS*, 2021.

D. Hendrycks and T. Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *ICLR*, 2019.

G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt. OpenCLIP. GitHub Repository, 2022.

D. D. Johnson, A. E. Hanchi, and C. J. Maddison. Contrastive Learning Can Find An Optimal Basis For Approximately View-Invariant Functions. In *ICLR*, 2023.

B. T. Kiani, R. Balestriero, Y. Chen, S. Lloyd, and Y. Le-Cun. Joint Embedding Self-Supervised Learning in the Kernel Regime. arXiv Technical Report, 2022.

I. Klebanov, I. Schuster, and T. J. Sullivan. A Rigorous Theory of Conditional Mean Embeddings. *SIAM Journal on Mathematics of Data Science*, 2020.

I. Klebanov, B. Sprungk, and T. Sullivan. The linear conditional expectation in Hilbert space. *Bernoulli*, 2021.

H. O. Lancaster. The Structure of Bivariate Distributions. *The Annals of Mathematical Statistics*, 1958.

H. Larochelle, D. Erhan, and Y. Bengio. Zero-data Learning of New Tasks. In *AAAI*, 2008.

S. L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

J. D. Lee, Q. Lei, N. Saunshi, and J. Zhuo. Predicting What You Already Know Helps: Provable Self-Supervised Learning. In *NeurIPS*, 2021.

A. Li, A. Jabri, A. Joulin, and L. van der Maaten. Learning Visual N-Grams from Web Data . In *ICCV*, 2017.

Y. Li, R. Pogodin, D. J. Sutherland, and A. Gretton. Self-Supervised Learning with Kernel Dependence Maximization. In *NeurIPS*, 2021.

Z. Li, D. Meunier, M. Mollenhauer, and A. Gretton. Towards Optimal Sobolev Norm Rates for the Vector-Valued Regularized Least-Squares Algorithm. *JMLR*, 2024.

Llama Team, Meta AI. The Llama 3 Herd of Models. arXiv Technical Report, 2024.

S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi. Fine-Grained Visual Classification of Aircraft. arXiv Technical Report, 2013.

M. Maniparambil, C. Vorster, D. Molloy, N. Murphy, K. McGuinness, and N. E. O'Connor. Enhancing CLIP with GPT-4: Harnessing Visual Descriptions as Prompts. In *ICCV*, 2023.

S. Menon and C. Vondrick. Visual Classification via Description from Large Language Models. In *ICLR*, 2023.

D. Meunier, Z. Shen, M. Mollenhauer, A. Gretton, and Z. Li. Optimal Rates for Vector-Valued Spectral Regularization Learning Algorithms. In *NeurIPS*, 2024.

T. Michaeli, W. Wang, and K. Livescu. Nonparametric Canonical Correlation Analysis. In *ICML*, 2016.

D. H. Nguyen, W. Zellinger, and S. Pereverzyev. On Regularized Radon-Nikodym Differentiation. *JMLR*, 2024.

M.-E. Nilsback and A. Zisserman. Automated Flower Classification over a Large Number of Classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

K. Oko, L. Lin, Y. Cai, and S. Mei. A Statistical Theory of Contrastive Pre-training and Multimodal Generative AI. arXiv Technical Report, 2025.

R. Parhi and R. D. Nowak. Banach Space Representer Theorems for Neural Networks and Ridge Splines. *JMLR*, 2021.

I. F. Pinelis and A. I. Sakhanenko. Remarks on Inequalities for Large Deviation Probabilities. *Theory of Probability & Its Applications*, 1986.

A. Pokle, J. Tian, Y. Li, and A. Risteski. Contrasting the landscape of contrastive and non-contrastive learning. In *AISTATS*, 2022.

S. Pratt, I. Covert, R. Liu, and A. Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *ICCV*, 2023.

R. Pryzant, D. Iter, J. Li, Y. Lee, C. Zhu, and M. Zeng. Automatic Prompt Optimization with "Gradient Descent" and Beam Search. In *EMNLP*, 2023.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do ImageNet Classifiers Generalize to ImageNet? In *ICML*, 2019.

A. Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 1959.

N. Saunshi, O. Plevrakis, S. Arora, M. Khodak, and H. Khandeparkar. A Theoretical Analysis of Contrastive Unsupervised Representation Learning. In *ICML*, 2019.

M. Scetbon and Z. Harchaoui. Harmonic Decompositions of Convolutional Networks. In *ICML*, 2020.

M. Scetbon, J. Jennings, A. Hilmkil, C. Zhang, and C. Ma. A fixed-point approach for causal generative modeling. In *ICML*, 2024.

R. Schilling. *Measures, Integrals, and Martingales*. Springer, 2nd edition, 2017.

J. Schmidt-Hieber. Rejoinder: Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 2020.

M. Sclar, Y. Choi, Y. Tsvetkov, and A. Suhr. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. In *ICLR*, 2024.

G. R. Shorack. *Probability for Statisticians*, volume 951. Springer, 2000.

R. Shwartz-Ziv, R. Balestriero, K. Kawaguchi, T. G. J. Rudner, and Y. LeCun. An Information Theory Perspective on Variance-Invariance-Covariance Regularization. In *NeurIPS*, 2023.

J. W. Siegel and J. Xu. Characterization of the Variation Spaces Corresponding to Shallow Neural Networks. *Constructive Approximation*, 2023.

S. Smale and D.-X. Zhou. Learning Theory Estimates via Integral Operators and Their Approximations. *Constructive Approximation*, 2007.

I. Steinwart and C. Scovel. Mercer's Theorem on General Domains: On the Interaction between Measures, Kernels, and RKHSs. *Constructive Approximation*, 2012.

Z. Tan, Y. Zhang, J. Yang, and Y. Yuan. Contrastive Learning is Spectral Clustering on Similarity Graph. In *ICLR*, 2024.

B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. YFCC100M: the New Data in Multimedia Research. *Communications of the ACM*, 2016.

S. Thrun and L. Pratt. *Learning to Learn*. Springer, 1998.

C. Tosh, A. Krishnamurthy, and D. Hsu. Contrastive learning, multi-view redundancy, and linear models. In *ALT*, 2021.

M. Unser. Ridges, Neural Networks, and the Radon Transform. *JMLR*, 2023.

A. Visheratin. NLLB-CLIP - train performant multilingual image retrieval model on a budget. In *NeurIPS Workshop: ENLSP-III*, 2023.

G. Wahba. *Spline models for observational data*. SIAM, 1990.

T. Wang and P. Isola. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. In *ICML*, 2020.

X. Wang, C. Li, Z. Wang, F. Bai, H. Luo, J. Zhang, N. Jojic, E. Xing, and Z. Hu. PromptAgent: Strategic Planning with Language Models Enables Expert-level Prompt Optimization. In *ICLR*, 2024.

Y. Wang and M. I. Jordan. Desiderata for Representation Learning: A Causal Perspective. *JMLR*, 2024.

Z. Wen and Y. Li. Toward Understanding the Feature Learning Process of Self-supervised Contrastive Learning. In *ICML*, 2021.

L. Wu and J. Long. A spectral-based analysis of the separation between two-layer neural networks and linear methods. *JMLR*, 2022.

J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

H. Xu, S. Xie, X. Tan, P.-Y. Huang, R. Howes, V. Sharma, S.-W. Li, G. Ghosh, L. Zettlemoyer, and C. Feichtenhofer. Demystifying CLIP data. In *ICLR*, 2024.

Y. Yang, A. Panagopoulou, S. Zhou, D. Jin, C. Callison-Burch, and M. Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *CVPR*, 2023.

J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In *ICML*, 2021.

J. Zhang, J. Jennings, A. Hilmkil, N. Pawlowski, C. Zhang, and C. Ma. Towards causal foundation model: on duality between optimal balancing and attention. In *ICML*, 2024.

Appendix

Table of Contents

A. Notation

Symbol	Description
$\boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathcal{Y}, \boldsymbol{z} \in \mathcal{Z}$	Instances and sample spaces for data modalities/views, often images, labels, and captions.
$\boldsymbol{\alpha}, \boldsymbol{\beta}$	Encoders $\boldsymbol{\alpha} : \mathcal{X} \to \mathbb{R}^d$ and $\boldsymbol{\beta} : \mathcal{Z} \to \mathbb{R}^d$.
(X, Y, Z)	Random variable realized in $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$.
$P_{X,Y}$	Evaluation distribution over $\mathcal{X} \times \mathcal{Y}$.
$Q_{X,Z}$	Pre-training distribution over $\mathcal{X} \times \mathcal{Z}$.
$\rho_{Y,Z}$	Prompting distribution over $\mathcal{Y} \times \mathcal{Z}$.
r	A function $r : \mathcal{Y} \to \mathbb{R}$.
$\eta_\star(\boldsymbol{x})$	Direct predictor $\mathbb{E}_{P_{X,Y}}\left[r(Y)|X\right](\boldsymbol{x})$.
$g_\rho(\boldsymbol{z})$	Prediction function $\mathbb{E}_{\rho_{Z,Y}}\left[\star(Y)|Z\right](\boldsymbol{z})$.
$\eta_\rho(\boldsymbol{x})$	Indirect predictor $\mathbb{E}_{Q_{X,Z}}\left[g_\rho(Z)|X\right](\boldsymbol{x})$.
N	Sample size of pre-training set $(X_1, Z_1), \ldots, (X_N, Z_N) \overset{\text{i.i.d}}{\sim} Q_{X,Z}$.
M	Number of prompts $(Y_1, Z_1) \ldots, (Y_M, Z_M) \overset{\text{i.i.d}}{\sim} \rho_{Y,Z}$.
$\mathbf{L}^2(P_X)$	Set containing equivalence classes of measurable functions $h : \mathcal{X} \to \mathbb{R}$ satisfying $\|h\|^2_{\mathbf{L}^2(P_X)} = \int h^2(\boldsymbol{x})\mathrm{d}P_X(\boldsymbol{x}) < +\infty$.
$\mathbf{M}_{Z|X}$	Conditional mean operator $[\mathbf{M}_{Z|X}g](\boldsymbol{z}) = \mathbb{E}_{Q_{X,Z}}\left[g(Z)|X\right](\boldsymbol{x})$.
R	Information density $\frac{\mathrm{d}Q_{X,Z}}{\mathrm{d}Q_X Q_Z} : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$.
$D_{\chi^2}(P\|Q)$	χ^2-divergence $\mathbb{E}_{U \sim Q}\left[\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(U) - 1\right)^2\right]$.
$I(X; Z)$	Mean square contingency $D_{\chi^2}(Q_{X,Z}\|Q_X Q_Z)$.
$(\sigma_i)_{i=1}^\infty$	Singular values of $\mathbf{M}_{Z|X}$.
$(\alpha_i, \beta_i)_{i=1}^\infty$	Left and right singular functions of $\mathbf{M}_{Z|X}$.
$\mathsf{S}_{\boldsymbol{z}}$	Conditional information density $\frac{\mathrm{d}P_{X,Y|\boldsymbol{z}}}{\mathrm{d}P_{X|\boldsymbol{z}} Q_{X|\boldsymbol{z}}} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$.
$I(X; Y|\boldsymbol{z})$	Conditional mean square contingency $D_{\chi^2}(P_{X,Y|\boldsymbol{z}}\|P_{X|\boldsymbol{z}} P_{Y|\boldsymbol{z}})$.
$\|\cdot\|_{\mathrm{HS}(\mathcal{G}, \mathcal{H})}$	Hilbert-Schmidt norm of a linear operator from $\mathcal{G}$ to $\mathcal{H}$.

Table 1. Notation used throughout the main text.

In the appendix, we use slightly more explicit notation. For example, the product measure of Q_X and Q_Z on $\mathcal{X} \times \mathcal{Z}$ is denoted $Q_X \otimes Q_Z$. The bracket notation $[\cdot]_X$ and $[\cdot]_Z$ are used to indicate equivalence classes in $\mathbf{L}^2(Q_X)$ and $\mathbf{L}^2(Q_Z)$, respectively. Such changes are marked as they are introduced.

B. Technical Background

In this section, we review the necessary background and construct any theoretical tools used in our analyses in a self-contained manner. Appx. B.1 describes the broadest function class we consider and gives a rigorous description of the conditional means that we employ in this work. Appx. B.2 reviews the basic classes of linear operators (trace class, Hilbert-Schmidt, etc.) that we consider. Appx. B.3 contains central tools regarding the structure of bivariate distributions. Finally, Appx. B.4 contains a brief introduction to reproducing kernel Hilbert spaces and some recent statistical results used in the proofs of Thm. 2 and Thm. 3.

B.1. Conditional Expectation and the Hilbert Space $\mathbf{L}^2$

Consider a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a topological space $\mathcal{X}$ equipped with its Borel σ-algebra $\mathcal{B}(\mathcal{X})$. Given a random variable $X : \Omega \to \mathcal{X}$ representing some observable data, we consider P_X to be the law of X, i.e. $P_X(B) = \mathbb{P}(X^{-1}(B))$ for every Borel set $B \in \mathcal{B}(\mathcal{X})$. Our goal is to define $\mathbf{L}^2(P_X)$, a Hilbert space containing equivalence classes of functions that are square integrable under P_X. As an intermediate step, we will also construct a Hilbert space $\mathsf{L}^2(\mathcal{G})$ for the σ-algebra $\mathcal{G} \subseteq \mathcal{F}$, which contains equivalence classes of $\mathcal{G}$-measurable functions that are square integrable under $\mathbb{P}$. Having both of these constructions will be helpful in working with conditional mean operators in a rigorous manner.

Quotient Space. As a starting point, consider the set

$$\mathsf{L}^2_+(\mathcal{F}) := \left\{ \mathcal{F}\text{-measurable functions } u : \Omega \to \mathbb{R} \text{ satisfying } \|u\|^2_{\mathsf{L}^2_+(\mathcal{F})} := \int_\Omega u^2(\omega) \mathrm{d}\mathbb{P}(\omega) < \infty \right\}.$$

For any $u, v \in \mathsf{L}^2_+(\mathcal{F})$, consider the equivalence relation "$\sim$" defined by

$$u \sim v \iff \exists \Omega_1 \in \mathcal{F} \text{ such that } u(\omega) = v(\omega) \ \forall \omega \in \Omega_1 \text{ and } \mathbb{P}(\Omega_1) = 1. \tag{17}$$

For any $u_+ \in \mathsf{L}^2_+(\mathcal{F})$, we define $[u_+]_\sim \in \mathsf{L}^2(\mathcal{F})$ as indexing the equivalence class containing all functions that differ from u_+ only on a set of $\mathbb{P}$-measure zero. The global Hilbert space will be defined using the quotient of $\mathsf{L}^2_+(\mathcal{F})$ under this equivalence relation.

Lemma 1. *The quotient space $\mathsf{L}^2(\mathcal{F}) = \mathsf{L}^2_+(\mathcal{F})/ \sim$ is a Hilbert space with the addition and scalar multiplication rules*

$$(u, v) \mapsto au + bv := [au_+ + bv_+]_\sim \text{ for some } u_+ \in u \text{ and } v_+ \in v,$$

for scalars $a, b \in \mathbb{R}$ and the inner product

$$(u, v) \mapsto \langle u, v \rangle_{\mathsf{L}^2(\mathcal{F})} := \int_\Omega u_+(\omega) v_+(\omega) \mathrm{d}\mathbb{P}(\omega) \text{ for some } u_+ \in u \text{ and } v_+ \in v,$$

where the definitions are independent of the choice of u_+ and v_+.

Proof. It is easy to verify that the addition, scalar multiplication, and inner product operations are well-defined (i.e. are invariant to the choice of u_+ and v_+). Define the norm $u \mapsto \|u\|_{\mathsf{L}^2(\mathcal{F})} := \sqrt{\langle u, u \rangle_{\mathsf{L}^2(\mathcal{F})}}$, and consider a Cauchy sequence $(u^{(n)})_{n=1}^\infty$ in $\mathsf{L}^2(\mathcal{F})$. To confirm completeness, we identify a limit of this sequence as an element of $\mathsf{L}^2(\mathcal{F})$. First, consider an arbitrary sequence $u_+^{(1)}, u_+^{(2)}, \dots$ where $u_+^{(n)} \in u^{(n)}$ for all $n \geq 1$. Then, we have by the Riesz-Fischer theorem (Schilling, 2017, Theorem 13.7), there exists a limit $u_+ \in \mathsf{L}^2_+(\mathcal{F})$ such that

$$\lim_{n \to \infty} \|u_+^{(n)} - u_+\|_{\mathsf{L}^2_+(\mathcal{F})} \to 0. \tag{18}$$

We then define $\lim_{n \to \infty} u^{(n)} := [u_+]_\sim$, and see that

$$\|u^{(n)} - [u_+]_\sim\|_{\mathsf{L}^2(\mathcal{F})} = \|u_+^{(n)} - u_+\|_{\mathsf{L}^2_+(\mathcal{F})} \to 0 \text{ as } n \to \infty,$$

where the last step follows by (18) and completes the proof. $\qquad\square$

Next, we construct closed subspaces of $\mathsf{L}^2(\mathcal{F})$ which contain random variables that are measurable functions of another

random variable. Notice that for $u, v \in \mathsf{L}^2(\mathcal{F})$, the statement $u = v$ indicates equality of two partitions, namely collections of random variables that differ pairwise on sets of measure zero. Letting $\sigma(X)$ denote the σ-algebra generated by X, define the set

$$\mathsf{L}^2_+(\sigma(X)) := \left\{ u \in \mathsf{L}^2_+(\mathcal{F}) \text{ s.t. } u \text{ is } \sigma(X)\text{-measurable} \right\}. \tag{19}$$

Then, using the equivalence relation (17), we define the space

$$\mathsf{L}^2(\sigma(X)) := \mathsf{L}^2_+(\sigma(X))/ \sim .$$

In the upcoming Cor. 1, we will confirm that $\mathsf{L}^2(\sigma(X))$ is indeed a closed subspace of $\mathsf{L}^2(\mathcal{F})$ for any random variable X. Before doing so, we consider the induced probability space $(\mathcal{X}, \mathcal{B}(\mathcal{X}), P_X)$. Then, we define the related linear space

$$\mathbf{L}^2_+(P_X) := \left\{ \text{measurable functions } f : \mathcal{X} \to \mathbb{R} \text{ satisfying } \|f\|^2_{\mathbf{L}^2_+(P_X)} := \int_{\mathcal{X}} f^2(\boldsymbol{x}) \mathrm{d}P_X(\boldsymbol{x}) < \infty \right\}.$$

We define an analogous equivalence relation "$\sim_X$" defined as

$$f \sim_X g \iff \exists \mathcal{X}_1 \in \mathcal{B}(\mathcal{X}) \text{ such that } f(\boldsymbol{x}) = g(\boldsymbol{x}) \, \forall \boldsymbol{x} \in \mathcal{X}_1 \text{ and } P_X(\mathcal{X}_1) = 1, \tag{20}$$

and the quotient $\mathbf{L}^2(P_X) := \mathbf{L}^2_+(P_X)/ \sim_X$. These sets are related to one another in the following lemma.

Corollary 1. *The set $\mathsf{L}^2(\sigma(X))$ is a Hilbert space with respect to the inner product used in Lem. 1, whereas $\mathbf{L}^2(P_X)$ is a Hilbert space with respect to the analogous inner product for $(\mathcal{X}, \mathcal{B}(\mathcal{X}), P_X)$. Furthermore, $\mathsf{L}^2(\sigma(X))$ is a closed subspace of $\mathsf{L}^2(\mathcal{F})$, and it holds that*

$$\mathsf{L}^2(\sigma(X)) = \mathbf{L}^2(P_X) \circ X := \left\{ [f_+(X(\cdot))]_\sim : f_+ \in \mathbf{L}^2_+(P_X) \right\}. \tag{21}$$

Proof. That $\mathsf{L}^2(\sigma(X))$ and $\mathbf{L}^2(P_X)$ are Hilbert spaces follows by identical arguments to Lem. 1. Additionally, we may invoke Schilling (2017, Lemma 27.1) to assert that $\mathsf{L}^2(\sigma(X))$ is a closed subspace of $\mathsf{L}^2(\mathcal{F})$. Finally, to show (21), we will show that

$$\mathsf{L}^2_+(\sigma(X)) = \mathbf{L}^2_+(P_X) \circ X := \left\{ f_+(X(\cdot)) : f_+ \in \mathbf{L}^2_+(P_X) \right\}$$

and take the quotient with respect to "$\sim$" on either side to complete the proof. First, $\mathbf{L}^2_+(P_X) \circ X \subseteq \mathsf{L}^2_+(\sigma(X))$ holds because $f_+(X)$ is clearly $\sigma(X)$-measurable and

$$\|f_+(X)\|^2_{\mathsf{L}^2_+(\mathcal{F})} = \int_\Omega (f_+(X(\omega)))^2 \mathrm{d}\mathbb{P}(\omega) \overset{P_X = X_\# \mathbb{P}}{=} \int_{\mathcal{X}} f^2_+(\boldsymbol{x}) \mathrm{d}P_X(\boldsymbol{x}) = \|f_+\|^2_{\mathbf{L}^2_+(P_X)} < \infty. \tag{22}$$

To show that $\mathsf{L}^2_+(\sigma(X)) \subseteq \mathbf{L}^2_+(P_X) \circ A$, first note that for any $\sigma(X)$-measurable random variable U, there exists a measurable function $g_+ : \mathcal{X} \to \mathbb{R}$ such that $U = g_+(X)$ (Durrett, 2019, Exercise 1.3.8). Applying (22) gives $\|g_+\|^2_{\mathbf{L}^2_+(P_X)} = +\infty \implies \|g_+(X)\|^2_{\mathsf{L}^2_+(\mathcal{F})} = +\infty$, which yields a contradiction as $\|g_+(X)\|^2_{\mathsf{L}^2_+(\mathcal{F})} = \|U\|^2_{\mathsf{L}^2_+(\mathcal{F})} < +\infty$. Thus, $\|g_+\|^2_{\mathbf{L}^2_+(P_X)} < +\infty$, completing the proof. $\qquad\square$

Conditional Expectation. Using Cor. 1, for any collection of random variables (X, Z, Y), we can now construct the Hilbert subspaces $\mathbf{L}^2(P_{X,Y})$, $\mathbf{L}^2(P_X)$, $\mathbf{L}^2(P_Z)$. We can then identify them with conditional expectations, i.e. projections onto $\mathsf{L}^2(\sigma(X, Y))$, $\mathsf{L}^2(\sigma(X))$, $\mathsf{L}^2(\sigma(Z))$, respectively. This is done in the definition below.

Definition 3 (Conditional Expectation). For any random variable $U \in \mathsf{L}^2(\mathcal{F})$, we define the *conditional expectation* $\mathbb{E}[U|\sigma(X)]$ as the orthogonal projection of U onto $\mathsf{L}^2(\sigma(X))$, or

$$\mathbb{E}[U|\sigma(X)] := \underset{u \in \mathsf{L}^2(\sigma(X))}{\arg\min} \|u - U\|^2_{\mathsf{L}^2(\mathcal{F})},$$

which uniquely exists due to the closedness of $\mathsf{L}^2(\sigma(X))$ and the projection theorem (Schilling, 2017, Theorem 26.13).

Owing to Cor. 1, we will also define the *conditional expectation function*

$$\mathbb{E}\left[U|X\right] : \mathcal{X} \to \mathbb{R}$$

as any measurable function satisfying the conditions $[\mathbb{E}\left[U|X\right]]_{\sim} \in \mathbf{L}^2(P_X)$ and $\mathbb{E}\left[U|\sigma(X)\right](\omega) = \mathbb{E}\left[U|X\right](X(\omega))$ for $\mathbb{P}$-almost every $\omega \in \Omega$. The specific function choice will not affect any of the forthcoming arguments.

Here, we defined the conditional expectation as an element of $\mathsf{L}^2(\sigma(X))$ and associated it with a function in $\mathbf{L}^2(P_X)$. Without the squared-integrability requirement, the conditional expectation may also be defined using the familiar tower property. We include the tower property below for completeness.

Lemma 2. *(Schilling, 2017, Theorem 27.12) Consider $U \in \mathsf{L}^2(\mathcal{F})$ and $X : \Omega \to \mathcal{X}$. Then, for every measurable set $A \in \sigma(X)$, it holds that*

$$\int_A U(\omega)\mathrm{d}\mathbb{P}(\omega) = \int_A \mathbb{E}\left[U|\sigma(X)\right](\omega)\mathrm{d}\mathbb{P}(\omega) = \int_{X(A)} \mathbb{E}\left[U|X\right](\boldsymbol{x})\mathrm{d}P_X(\boldsymbol{x}).$$

We will make use of both the projection property and tower property throughout this manuscript. While conditional expectation may be defined for specific integrable functions, we may wish to define probability measures whose integrals can produce all conditional expectations simultaneously—this ideal is captured by *regular conditional distributions (r.c.d.'s)* (Shorack, 2000), which we recall below.

Definition 4. Consider random variables $(U, V) : \Omega \to \mathcal{U} \times \mathcal{V}$. Let $\mathcal{B}(\mathcal{U})$ denote the Borel σ-algebra on $\mathcal{U}$. A map: $\mu : \mathcal{V} \times \mathcal{B}(\mathcal{U}) :\to [0, 1]$ is called a *regular conditional distribution (r.c.d.)* if the following two properties hold:

1. For each $A \in \mathcal{B}(\mathcal{U})$ and $\boldsymbol{v} \in \mathcal{V}$, it holds that

   $$\mu(\boldsymbol{v}, A) = \mathbb{E}_{P_{U,V}}\left[\mathbb{1}_A(U)|V\right](\boldsymbol{v}),$$

   for the conditional expectation defined in Definition 3.

2. For P_V-almost every $\boldsymbol{v} \in \mathcal{V}$, $\mu(\boldsymbol{v}, \cdot)$ is a probability measure on $\mathcal{B}(\mathcal{U})$.

This will primarily be used for the conditional dependence arguments in Appx. C.

B.2. Compact Operators

We collect several generalities about Hilbert spaces and linear operators (hereafter, simply "operators") between them. Many computations will require expanding an element of a separable Hilbert space onto an orthonormal basis.

Definition 5 (Separability, Orthonormal Basis, Complete Orthonormal System). For a Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ over $\mathbb{R}$, the orthonormal system $e_1, e_2, \ldots \in \mathcal{H}$ of vectors is called an *orthonormal basis (ONB)* or *complete orthonormal system (CONS)* of $\mathcal{H}$ if any of the following properties hold, which are equivalent by Schilling (2017, Theorem 26.21).

1. For every $h \in \mathcal{H}$, $\langle h, e_i \rangle_{\mathcal{H}} = 0$ for all $i \geq 1$ implies that $h \equiv 0$.

2. $\bigcup_{n=1}^{\infty} \operatorname{span}\{e_1, \ldots, e_n\}$ is dense in $\mathcal{H}$.

3. For every $h \in \mathcal{H}$, it holds that $h = \sum_{i=1}^{\infty} \langle h, e_i \rangle_{\mathcal{H}} e_i$.

4. For every $h \in \mathcal{H}$, it holds that $\sum_{i=1}^{\infty} |\langle h, e_i \rangle_{\mathcal{H}}|^2 = \|h\|_{\mathcal{H}}^2$.

5. For every $h, h' \in \mathcal{H}$, it holds that $\sum_{i=1}^{\infty} \langle h, e_i \rangle_{\mathcal{H}} \langle h', e_i \rangle_{\mathcal{H}} = \langle h, h' \rangle_{\mathcal{H}}$.

If there exists a countable orthonormal basis, then $\mathcal{H}$ is called *separable* (Schilling, 2017, Definition 26.23 & Theorem 26.24).

When linear operators are compact, then we may decompose them in a way that generalizes the eigendecomposition and singular value decomposition for matrices.

Definition 6 (Compact Operator). A linear operator $\mathbf{M} : \mathcal{G} \to \mathcal{H}$ between Hilbert spaces $\mathcal{G}$ and $\mathcal{H}$ is called *compact* if for every totally bounded subset $B \subseteq \mathcal{G}$, the image $\mathbf{M}(B)$ is relatively compact (i.e. the closure of $\mathbf{M}(B)$ is compact) in $\mathcal{H}$.

Compact operators are bounded, and every bounded linear operator $\mathbf{M}$ admits a unique *adjoint operator* $\mathbf{M}^*$ satisfying $\langle h, \mathbf{M}g \rangle_{\mathcal{H}} = \langle \mathbf{M}^*h, g \rangle_{\mathcal{G}}$ for all $g \in \mathcal{G}$ and $h \in \mathcal{H}$. An operator $\mathbf{T} : \mathcal{H} \to \mathcal{H}$ is called *self-adjoint* if $\mathbf{T} = \mathbf{T}^*$. Next, we collect two operator decompositions that will be used repeatedly. We refer the reader to Gohberg et al. (2003, Chapter IV) and Gohberg et al. (2003, Chapter X) for further discussion on these topics. Just as their analogs for matrices, we refer to them as the *eigendecomposition* and *singular value decomposition*, respectively.

Theorem 4. *(Gohberg et al., 2003, Chapter IV, Theorem 5.1) Let $\mathbf{T} : \mathcal{H} \to \mathcal{H}$ be a compact, self-adjoint operator on a separable Hilbert space $\mathcal{H}$ on $\mathbb{R}$. Then, there exists a countable orthonormal basis $\{e_j\}_{j \in J}$ of $\mathcal{H}$ and a sequence of non-zero real numbers $\{\lambda_i\}_{i \in I}$ with $\lambda_i \to 0$, $I \subseteq J$, and for all $h \in \mathcal{H}$, we have that*

$$\mathbf{T}h = \sum_{i \in I} \lambda_i \langle h, e_i \rangle_{\mathcal{H}} e_i. \tag{23}$$

Furthermore, if $\langle h, \mathbf{T}h \rangle_{\mathcal{H}} \geq 0$ for all $h \in \mathcal{H}$ (i.e. $\mathbf{T}$ is positive semidefinite*), then we may take $\lambda_i > 0$ for all $i \in I$, and order them in a non-increasing sequence.. We call $\{\lambda_i\}_{i \in I}$ the non-zero* eigenvalues *of $\mathbf{T}$.*

Theorem 5. *(Gohberg et al., 2003, Chapter X, Theorem 4.2) Let $\mathbf{M} : \mathcal{G} \to \mathcal{H}$ be a compact operator between separable Hilbert spaces $\mathcal{G}$ and $\mathcal{H}$ on $\mathbb{R}$. Then, there exists an orthonormal basis $\{u_j\}_{j \in J}$ of $\mathcal{H}$, an orthonormal basis $\{v_k\}_{k \in K}$ of $\mathcal{G}$, and a sequence of positive real numbers $\{s_i\}_{i \in I}$ with $s_i \to 0$ such that the following statements hold.*

- *All collections are at most countable, i.e. $I, J, K \subseteq \mathbb{N}$, and $I \subseteq J \cap K$.*

- *For all $g \in \mathcal{G}$ and $h \in \mathcal{H}$, we have that*

$$\mathbf{M}g = \sum_{i \in I} s_i \langle g, v_i \rangle_{\mathcal{G}} u_i \text{ and } \mathbf{M}^*h = \sum_{i \in I} s_i \langle h, u_i \rangle_{\mathcal{H}} v_i. \tag{24}$$

We call $\{s_i\}_{i \in I}$ the non-zero singular values *of $\mathbf{M}$, which can be ordered in a non-increasing sequence.*

The sets J and K are used to index the bases of $\mathcal{H}$ and $\mathcal{G}$, so they may be larger in cardinality than I, which only indexes the non-zero eigenvalue and singular values, respectively. We will also consider more specific classes of compact operators.

Definition 7. A compact operator $\mathbf{M}$ with singular values $\{s_i\}_{i \in I}$ (Thm. 5) is called *trace class* if $\sum_{i \in I} s_i < +\infty$ (the singular values are summable) and *Hilbert-Schmidt* if $\sum_{i \in I} s_i^2 < +\infty$ (the singular values are square summable).

Using the singular value decomposition, we see that if $\mathbf{M}$ is Hilbert-Schmidt, then $\mathbf{M}\mathbf{M}^*$ and $\mathbf{M}^*\mathbf{M}$ are self-adjoint trace class operators. The set of all Hilbert-Schmidt operators $\mathbf{M} : \mathcal{G} \to \mathcal{H}$ between Hilbert spaces $(\mathcal{G}, \langle \cdot, \cdot \rangle_{\mathcal{G}})$ and $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ will be denoted by $\mathrm{HS}(\mathcal{G}, \mathcal{H})$. This is itself a Hilbert space with the inner product

$$\langle \mathbf{A}, \mathbf{B} \rangle_{\mathrm{HS}(\mathcal{G}, \mathcal{H})} = \sum_{j \in J} \langle \mathbf{A}g_j, \mathbf{B}g_j \rangle_{\mathcal{H}}$$

where $\{g_j\}_{j \in J}$ can be taken to be *any* orthonormal basis of $\mathcal{G}$. Similarly, let $\{h_k\}_{k \in K}$ be an arbitrary orthonormal basis of $\mathcal{H}$. Then, the Hilbert-Schmidt norm $\|\mathbf{A}\|_{\mathrm{HS}(\mathcal{G}, \mathcal{H})}$ will be defined as

$$
\begin{aligned}
\|\mathbf{A}\|_{\mathrm{HS}(\mathcal{G}, \mathcal{H})}^2 &= \langle \mathbf{A}, \mathbf{A} \rangle_{\mathrm{HS}(\mathcal{G}, \mathcal{H})} \\
&= \sum_{j \in J} \langle \mathbf{A}g_j, \mathbf{A}g_j \rangle_{\mathcal{H}} \\
&= \sum_{j \in J} \sum_{k \in K} \sum_{l \in K} \langle \mathbf{A}g_j, h_k \rangle_{\mathcal{H}} \langle \mathbf{A}g_j, h_l \rangle_{\mathcal{H}} \langle h_k, h_l \rangle_{\mathcal{H}} \\
&= \sum_{j \in J} \sum_{k \in K} \langle h_k, \mathbf{A}g_j \rangle_{\mathcal{H}}^2 = \sum_{j \in J} \sum_{k \in K} \langle \mathbf{A}^*h_k, g_j \rangle_{\mathcal{G}}^2.
\end{aligned}
\tag{25}
$$

Using the singular value decomposition, we see that (25) is equal to the sum of the squared singular values referenced in Definition 7. For $h \in \mathcal{H}$ and $g \in \mathcal{G}$, we define the rank-one operator $h \otimes g : \mathcal{G} \to \mathcal{H}$ via $(h \otimes g)g' = \langle g, g' \rangle_{\mathcal{G}} h$ for all $g' \in \mathcal{G}$. For an operator $\mathbf{A} \in \mathrm{HS}(\mathcal{G}, \mathcal{H})$, the following identity regarding rank-one operators will be useful for norm computations:

$$\langle h, \mathbf{A}g \rangle_{\mathcal{H}} = \langle \mathbf{A}^* h, g \rangle_{\mathcal{G}} = \langle \mathbf{A}, h \otimes g \rangle_{\mathrm{HS}(\mathcal{G}, \mathcal{H})} = \langle \mathbf{A}^*, g \otimes h \rangle_{\mathrm{HS}(\mathcal{H}, \mathcal{G})}.$$

Finally, we will often compute Hilbert-Schmidt norms using assumptions on the singular decays of the operator in question.

Lemma 3. *Let $\mathbf{M} : \mathcal{G} \to \mathcal{H}$ be a Hilbert-Schmidt operator with singular values $\{s_i\}_{i \in I}$ (Thm. 5). Assume that $I = \mathbb{N}$ and that there exist constants $c, C, \gamma > 0$ such that $ci^{-\gamma} \leq s_i \leq Ci^{-\gamma}$ for all $i \in \mathbb{N}$. Then, $\gamma > 1/2$, and it holds that*

$$\frac{c^2}{2\gamma - 1} \leq \|\mathbf{M}\|_{\mathrm{HS}(\mathcal{G}, \mathcal{H})}^2 \leq \frac{2\gamma C^2}{2\gamma - 1}.$$

Proof. The requirement that $\gamma > 1/2$ follows from the square summability of $\{s_i\}_{i=1}^{\infty}$ and the bound $s_i \geq ci^{-\gamma}$. For the upper bound, write

$$\sum_{i=1}^{\infty} s_i^2 \leq C^2 \sum_{i=1}^{\infty} i^{-2\gamma} = C^2 \sum_{i=1}^{\infty} \int_{i-1}^{i} \lceil x \rceil^{-2\gamma} \, \mathrm{d}x \leq C^2 \left(1 + \int_1^{\infty} x^{-2\gamma} \mathrm{d}x \right) = \frac{2\gamma C^2}{2\gamma - 1}.$$

For the lower bound, write

$$\sum_{i=1}^{\infty} s_i^2 \geq c^2 \sum_{i=1}^{\infty} i^{-2\gamma} = c^2 \sum_{i=1}^{\infty} \int_i^{i+1} \lfloor x \rfloor^{-2\gamma} \, \mathrm{d}x \geq c^2 \int_1^{\infty} x^{-2\gamma} \mathrm{d}x = \frac{c^2}{2\gamma - 1},$$

the result as desired. $\qquad \square$

B.3. The Conditional Mean Operator

This section contains key properties of the conditional mean operator $\mathbf{M}_{Z|X}$ and the information density R from Sec. 2, based on the foundations of Appx. B.1 and Appx. B.2. As we shall show, owing to a particular *Lancaster decomposition* (Prop. 2), both operators enjoy convenient spectral representations and relate to a measure of dependence—the mean-squared contingency.

Recall the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Consider Borel measurable spaces $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ and $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$, and a random variable $(X, Z) : \Omega \to \mathcal{X} \times \mathcal{Z}$. We denote by $Q_{X,Z}$ the law of (X, Z), i.e. $Q_{X,Z}(B) = \mathbb{P}\left((X, Z)^{-1}(B) \right)$ for every $B \in \mathcal{B}(\mathcal{X} \times \mathcal{Z})$. Note that by Schilling (2017, Corollary 27.24), the Hilbert spaces $\mathbf{L}^2(Q_X)$, $\mathbf{L}^2(Q_Z)$, $\mathbf{L}^2(Q_{X,Z})$, and $\mathbf{L}^2(Q_X \otimes Q_Z)$ are separable, a fact we will maintain in this section. We use the notation $[\cdot]_X$ and $[\cdot]_Z$ to index equivalence classes in $\mathbf{L}^2(Q_X)$ and $\mathbf{L}^2(Q_Z)$, respectively. In other words, for a measurable function $h : \mathcal{X} \to \mathbb{R}$, we will write $[h]_X \in \mathbf{L}^2(Q_X)$ to indicate that $\int_{\mathcal{X}} h^2(\boldsymbol{x}) \mathrm{d}Q_X(\boldsymbol{x}) < +\infty$. Recall the conditional mean function introduced in Definition 3. We define the conditional mean operator

$$\begin{aligned} \mathbf{M}_{Z|X} &: \mathbf{L}^2(Q_Z) \to \mathbf{L}^2(Q_X) \\ \mathbf{M}_{Z|X}[g]_Z &= [\mathbb{E}_{Q_{X,Z}} [g(Z)|X] (\cdot)]_X. \end{aligned} \tag{26}$$

The specific function $g \in [g]_Z$ chosen for the output conditional expectation is not relevant, as all choices will result in the same equivalence class. We define $\mathbf{M}_{X|Z}$ as the analogous operator for the conditional mean of $h(X)$ given Z for $[h]_X \in \mathbf{L}^2(Q_X)$.

Spectral Representation. In the case that $\mathbf{M}_{Z|X}$ is compact, the conditional mean operator admits a singular value decomposition, which will be instrumental in obtaining several important properties.

Proposition 1 (Singular Value Decomposition of the Conditional Mean Operator). *Let $\mathbf{M}_{Z|X} : \mathbf{L}^2(Q_Z) \to \mathbf{L}^2(Q_X)$ be compact. There exists a countable orthonormal basis $\{\alpha_j\}_{j \in J}$ of $\mathbf{L}^2(Q_X)$, a countable orthonormal basis $\{\beta_k\}_{k \in K}$ of $\mathbf{L}^2(Q_Z)$, and a countable sequence of positive real numbers $\{\sigma_i\}_{i \in I}$ satisfying $\sigma_i \searrow 0$, $I \subseteq J \cap K$, and the following statements in addition:*

- *We may take $\sigma_1 = 1$, $\mathbf{1}_{\mathcal{X}} \in \alpha_1$, and $\mathbf{1}_{\mathcal{Z}} \in \beta_1$, where $\mathbf{1}_{\mathcal{X}}$ is identically 1 on $\mathcal{X}$ and $\mathbf{1}_{\mathcal{Z}}$ is identically 1 on $\mathcal{Z}$.*

- *For all $[g]_Z \in \mathbf{L}^2(Q_Z)$ and $[h]_X \in \mathbf{L}^2(Q_X)$, we have that*

$$\mathbf{M}_{Z|X}[g]_Z = \sum_{i \in I} \sigma_i \langle [g]_Z, \beta_i \rangle_{\mathbf{L}^2(Q_Z)} \alpha_i \text{ and } \mathbf{M}_{X|Z}[h]_X = \sum_{i \in I} \sigma_i \langle [h]_X, \alpha_i \rangle_{\mathbf{L}^2(Q_X)} \beta_i. \tag{27}$$

Proof. Beyond the direct application of Thm. 5, we must prove the statement regarding $(\sigma_1, \alpha_1, \beta_1)$ and that $\mathbf{M}_{Z|X}^* = \mathbf{M}_{X|Z}$ (which relates (24) to (27)) to achieve the desired result. For the first, we appeal to the variational representation of the first singular value σ_1 (Gohberg et al., 1990, Section IV.1, Eq. (2)), which states that

$$\sigma_1 = \sup \left\{ \|\mathbf{M}_{Z|X} g\|_{\mathbf{L}^2(Q_X)} : g \in \mathbf{L}^2(Q_Z), \|g\|_{\mathbf{L}^2(Q_Z)} = 1 \right\}. \tag{28}$$

We will show that $\beta_1 = [\mathbf{1}_{\mathcal{Z}}]_Z$ achieves the supremum. Then, it will hold that $\sigma_1 = 1$ and $\alpha_1 = \mathbf{M}_{Z|X}\beta_1 = [\mathbf{1}_{\mathcal{X}}]_X$, because any version of the conditional expectation $\mathbb{E}_{Q_{X,z}}[1|X]$ is Q_X-almost surely equal to 1. Consider any $[g]_Z \in \mathbf{L}^2(Q_Z)$ satisfying $\|g\|_{\mathbf{L}^2(Q_Z)} = 1$. Then, we apply Jensen's inequality and the tower property (Lem. 2) to achieve

$$\|\mathbf{M}_{Z|X}[g]_Z\|_{\mathbf{L}^2(Q_X)}^2 = \int_{\mathcal{X}} \left(\mathbb{E}_{Q_{X,z}}[g(Z)|X](\boldsymbol{x}) \right)^2 \mathrm{d}Q_X(\boldsymbol{x})$$

$$\leq \int_{\mathcal{X}} \mathbb{E}_{Q_{X,z}} \left[g^2(Z)|X \right](\boldsymbol{x}) \mathrm{d}Q_X(\boldsymbol{x})$$

$$= \mathbb{E}_{Q_Z} \left[g^2(Z) \right] = \|g\|_{\mathbf{L}^2(Q_Z)}^2 = 1.$$

Setting $g(\boldsymbol{z}) = \mathbf{1}_{\mathcal{Z}}(\boldsymbol{z}) \equiv 1$ achieves the upper bound, hence also achieving the supremum in (28). Next, to prove that $\mathbf{M}_{Z|X}^* = \mathbf{M}_{X|Z}$, we similarly consider $[h]_X \in \mathbf{L}^2(Q_X)$ and write

$$\langle [h]_X, \mathbf{M}_{Z|X}[g]_Z \rangle_{\mathbf{L}^2(Q_X)} = \mathbb{E}_{Q_X} \left[h(X)\mathbb{E}_{Q_{X,z}}[g(Z)|X] \right]$$

$$= \mathbb{E}_{Q_{X,z}}[h(X)g(Z)]$$

$$= \mathbb{E}_{Q_Z} \left[\mathbb{E}_{Q_{X,z}}[h(X)|Z] g(Z) \right]$$

$$= \langle \mathbf{M}_{X|Z}[h]_X, [g]_Z \rangle_{\mathbf{L}^2(Q_Z)},$$

which satisfies the adjoint relationship and completes the proof. $\square$

Lancaster Decomposition. In the remaining proofs of this section, we do not differentiate an equivalence class in an $\mathbf{L}^2$-space with its component functions, as the distinction will be clear from context. First, using the orthonormal bases defined in Prop. 1, we may form a convenient orthonormal basis of $\mathbf{L}^2(Q_X \otimes Q_Z)$.

Lemma 4. *The collection $\{\alpha_j \beta_k\}_{j \in J, k \in K}$ from Prop. 1, where $\{\alpha_j\}_{j \in J}$ is a countable orthonormal basis of $\mathbf{L}^2(Q_X)$ and $\{\beta_k\}_{k \in K}$ is a countable orthonormal basis of $\mathbf{L}^2(Q_Z)$, forms an orthonormal basis of $\mathbf{L}^2(Q_X \otimes Q_Z)$.*

Proof. We first show that $\{\alpha_j \beta_k\}_{j \in J, k \in K}$ is an orthonormal system. For any indices $i, i' \in I$ and $j, j' \in J$, it holds via independence that

$$\langle \alpha_j \beta_j, \alpha_{j'} \beta_{k'} \rangle_{\mathbf{L}^2(Q_X \otimes Q_Z)} = \langle \alpha_j, \alpha_{j'} \rangle_{\mathbf{L}^2(Q_X)} \langle \beta_k, \beta_{k'} \rangle_{\mathbf{L}^2(Q_Z)}$$

$$= \begin{cases} 1 & \text{if } j = j' \text{ and } k = k' \\ 0 & \text{otherwise} \end{cases}.$$

To establish that this orthonormal system is now complete, we use the first equivalent condition in Definition 5. Consider $s \in \mathbf{L}^2(Q_X \otimes Q_Z)$ such that $\langle s, \alpha_j \beta_k \rangle_{\mathbf{L}^2(Q_X \otimes Q_Z)} = 0$ for all $j \in J$ and $k \in K$. Then, via Fubini's theorem (Schilling,

2017, Corollary 14.9), it holds that

$$0 = \int_{\mathcal{Z}} \underbrace{\left(\int_{\mathcal{X}} s(\boldsymbol{x}, \boldsymbol{z}) \alpha_j(\boldsymbol{x}) \mathrm{d} Q_X(\boldsymbol{x}) \right)}_{g_j(\boldsymbol{z})} \beta_k(\boldsymbol{z}) \mathrm{d} Q_Z(\boldsymbol{z}) = \langle g_j, \beta_k \rangle_{\mathbf{L}^2(Q_Z)}.$$

Because $\{\beta_k\}_{k \in K}$ forms an ONB, it holds that the equivalence class of g_j is the zero element in $\mathbf{L}^2(Q_Z)$, or in other words, $g_j(\boldsymbol{z}) = 0$ for Q_Z-almost all $\boldsymbol{z} \in \mathcal{Z}$. Due to the fact that J is countable, we have that

$$\mathcal{Z}_1 := \bigcap_{j \in J} \{\boldsymbol{z} \in \mathcal{Z} : g_j(\boldsymbol{z}) = 0\} = \{\boldsymbol{z} \in \mathcal{Z} : g_j(\boldsymbol{z}) = 0 \ \forall j \in J\}$$

is a probability one set under Q_Z. Because $\{\alpha_j\}_{j \in J}$ is an ONB of $\mathbf{L}^2(Q_X)$, it also holds that

$$\{\boldsymbol{z} \in \mathcal{Z} : g_j(\boldsymbol{z}) = 0 \ \forall j \in J\} \subseteq \mathcal{Z}_1' = \{\boldsymbol{z} \in \mathcal{Z} : s(\boldsymbol{x}, \boldsymbol{z}) = 0 \text{ for } Q_X\text{-almost all } \boldsymbol{x} \in \mathcal{X}\},$$

indicating that the right-hand side is also a probability one set under Q_Z. Then, applying again the iterated integral,

$$\begin{aligned}
\int_{\mathcal{X} \times \mathcal{Z}} s^2(\boldsymbol{x}, \boldsymbol{z}) \mathrm{d}(Q_X \otimes Q_Z)(\boldsymbol{x}, \boldsymbol{z}) &= \int_{\mathcal{Z}} \left(\int_{\mathcal{X}} s^2(\boldsymbol{x}, \boldsymbol{z}) \mathrm{d} Q_X(\boldsymbol{x}) \right) \mathrm{d} Q_Z(\boldsymbol{z}) \\
&= \int_{\mathcal{Z}_1'} \left(\int_{\mathcal{X}} s^2(\boldsymbol{x}, \boldsymbol{z}) \mathrm{d} Q_X(\boldsymbol{x}) \right) \mathrm{d} Q_Z(\boldsymbol{z}) \\
&= 0,
\end{aligned}$$

indicating the $s(\boldsymbol{x}, \boldsymbol{z}) = 0$ for $(Q_X \otimes Q_Z)$-almost all $(\boldsymbol{x}, \boldsymbol{z}) \in \mathcal{X} \times \mathcal{Z}$. This completes the proof. $\qquad\square$

This basis allows us to relate the conditional mean operator to a particular Radon-Nikodym derivative. As a result, both can be used to measure the dependence between X and Z (Lancaster, 1958).

Proposition 2 (Lancaster Decomposition). *Assume that* $Q_{X,Z} \ll Q_X \otimes Q_Z$, *in which case there exists a Radon-Nikodym derivative* $\mathsf{R} = \frac{\mathrm{d} Q_{X,Z}}{\mathrm{d}(Q_X \otimes Q_Z)}$. *Then, the following identity holds:*

$$\mathsf{R} = \sum_{i \in I} \sigma_i \alpha_i \beta_i. \tag{29}$$

In particular, the operator $\mathbf{M}_{Z|X}$ *is Hilbert-Schmidt if and only if* $\mathsf{R} \in \mathbf{L}^2(Q_X \otimes Q_Z)$, *with the equality*

$$\|\mathbf{M}_{Z|X}\|^2_{\mathrm{HS}(\mathbf{L}^2(Q_Z), \mathbf{L}^2(Q_X))} = \|\mathsf{R}\|^2_{\mathbf{L}^2(Q_X \otimes Q_Z)} = \sum_{i \in I} \sigma_i^2.$$

Proof. Using Lem. 4, we represent R on the ONB $\{\alpha_j \beta_k\}_{j \in J, k \in K}$. For any $j \in J$ and $k \in K$, use the definition of the Radon-Nikodym derivative (Schilling, 2017, Theorem 20.2) to write

$$\begin{aligned}
\langle \mathsf{R}, \alpha_j \beta_k \rangle_{\mathbf{L}^2(Q_X \otimes Q_Z)} &= \mathbb{E}_{Q_X \otimes Q_Z} \left[\mathsf{R}(X, Z) \alpha_j(X) \beta_k(Z) \right] \\
&= \mathbb{E}_{Q_{X,Z}} \left[\alpha_j(X) \beta_k(Z) \right] \\
&= \mathbb{E}_{Q_X} \left[\alpha_j(X) \mathbb{E}_{Q_{X,Z}} \left[\beta_k(Z) | X \right] \right] \\
&= \langle \alpha_j, \mathbf{M}_{Z|X} \beta_k \rangle_{\mathbf{L}^2(Q_X)} \\
&= \begin{cases} \sigma_i & \text{if } j = k = i \in I \\ 0 & \text{otherwise} \end{cases},
\end{aligned}$$

where we recall I as the set indexing the non-zero singular values of $\mathbf{M}_{Z|X}$ (see Prop. 1). This proves (29), the first claim.

For the second claim, we use the orthonormality of $\{\alpha_j\beta_k\}_{j\in J, k\in K}$ in $\mathbf{L}^2(Q_X\otimes Q_Z)$, so that

$$\|\mathsf{R}\|^2_{\mathbf{L}^2(Q_X\otimes Q_Z)} = \sum_{i\in I}\sigma_i^2 = \|\mathbf{M}_{Z|X}\|_{\mathrm{HS}(\mathbf{L}^2(Q_Z),\mathbf{L}^2(Q_X))},$$

so that square-summability of $\{\sigma_i\}_{i\in I}$ implies finiteness and equality of the left-hand and right-hand sides above. $\qquad\square$

The formulas in Sec. 2 simply equated $I = \mathbb{N} = \{1, 2, \ldots\}$ for ease of presentation. For completeness, the ε_d term in (9) represents the tail of (29), i.e.,

$$\varepsilon_d = \sum_{i=d+1}^\infty \sigma_i\alpha_i\beta_i,$$

which vanishes as $d \to \infty$ because $\sigma_i \to 0$ and $\alpha_i\beta_i$ is unit-norm in $\mathbf{L}^2(Q_X\otimes Q_Z)$.

The Radon-Nikodym derivative $\mathsf{R} = \frac{\mathrm{d}Q_{X,Z}}{\mathrm{d}(Q_X\otimes Q_Z)}$ is also useful for converting conditional expectation computations into marginal expectation computations. In this sense, we may say that R acts as a kernel for an integral operator representation of $\mathbf{M}_{Z|X}$, where the integral is taken with respect to Q_Z. The following identity is referenced by Buja (1990, Section 3) and Dytso et al. (2023, Lemma 1, Eq. (14)). We provide a self-contained proof below.

Lemma 5. *Adopt the setting of Prop. 2. Then, for all $g \in \mathbf{L}^2(Q_Z)$ and $h \in \mathbf{L}^2(Q_X)$, it holds that*

$$\mathbb{E}_{Q_{X,Z}}[g(Z)|X](\boldsymbol{x}) = \mathbb{E}_{Q_Z}[g(Z)\mathsf{R}(\boldsymbol{x}, Z)] \text{ for } Q_X\text{-almost all } \boldsymbol{x} \in \mathfrak{X},$$
$$\mathbb{E}_{Q_{X,Z}}[h(X)|Z](\boldsymbol{z}) = \mathbb{E}_{Q_X}[h(X)\mathsf{R}(X, \boldsymbol{z})] \text{ for } Q_Z\text{-almost all } \boldsymbol{z} \in \mathfrak{Z}.$$

Proof. We prove the first identity, whereas the second follows by a symmetric argument. To confirm that the two functions are equal almost surely, it is sufficient to prove that for any measurable set $A \in \sigma(X)$ (the σ-algebra generated by X), the relation

$$\int_A \mathbb{E}_{Q_{X,Z}}[g(Z)|X](\boldsymbol{x})\mathrm{d}Q_X(\boldsymbol{x}) = \int_A \mathbb{E}_{Q_Z}[g(Z)\mathsf{R}(\boldsymbol{x}, Z)]\,\mathrm{d}Q_X(\boldsymbol{x}). \tag{30}$$

By the definition of conditional expectation, we have that

$$\int_A \mathbb{E}_{Q_{X,Z}}[g(Z)|X](\boldsymbol{x})\mathrm{d}Q_X(\boldsymbol{x}) = \int_{\mathfrak{X}} \mathbb{E}_{Q_{X,Z}}[g(Z)|X](\boldsymbol{x})\mathbb{1}_A(\boldsymbol{x})\mathrm{d}Q_X(\boldsymbol{x})$$
$$= \mathbb{E}_{Q_{X,Z}}[g(Z)\mathbb{1}_A(X)]$$
$$= \mathbb{E}_{Q_X\otimes Q_Z}[g(Z)\mathbb{1}_A(X)\mathsf{R}(X, Z)],$$

where the last step follows from the Radon-Nikodym theorem (Schilling, 2017, Theorem 20.2). Next, we compute the expectation, taken under the product measure, using Fubini's theorem (Schilling, 2017, Corollary 14.9). That is,

$$\int_A \mathbb{E}_{Q_{X,Z}}[g(Z)|X](\boldsymbol{x})\mathrm{d}Q_X(\boldsymbol{x}) = \mathbb{E}_{Q_X\otimes Q_Z}[g(Z)\mathbb{1}_A(X)\mathsf{R}(X, Z)]$$
$$= \int_A\left(\int_{\mathfrak{Z}} g(\boldsymbol{z})\mathsf{R}(\boldsymbol{x}, \boldsymbol{z})\mathrm{d}Q_Z(\boldsymbol{z})\right)\mathrm{d}Q_X(\boldsymbol{x})$$
$$= \int_A \mathbb{E}_{Q_Z}[g(Z)\mathsf{R}(\boldsymbol{x}, Z)]\,\mathrm{d}Q_X(\boldsymbol{x}).$$

This achieves (30) and completes the proof. $\qquad\square$

While Lem. 5 applies for a general function g, the function g_ρ from (5) is itself a conditional mean. This can be leveraged to produce yet another identity, which acts as a technical lemma for the proof of Thm. 3.

Lemma 6. *Recall that $g_\rho(\boldsymbol{z}) := \mathbb{E}_{\rho_{Y,Z}}[r(Y)|Z](\boldsymbol{z})$ for $r \in \mathbf{L}^2(P_Y)$. Assume in addition that $r \in \mathbf{L}^2(\rho_Y)$. Then,*

$$\eta_\rho(\boldsymbol{x}) = \mathbb{E}_{\rho_{Y,Z}}[r(Y)\mathsf{R}(\boldsymbol{x}, Z)] + \int_{\mathcal{Z}} g_\rho(\boldsymbol{z})\mathsf{R}(\boldsymbol{x}, \boldsymbol{z})(\mathrm{d}Q_Z(\boldsymbol{z}) - \mathrm{d}\rho_Z(\boldsymbol{z})).$$

for Q_X almost all $\boldsymbol{x} \in \mathfrak{X}$.

Proof. By Lem. 5, we already have that for Q_X-almost all $\boldsymbol{x} \in \mathfrak{X}$, the identity

$$\begin{aligned}
\eta_\rho(\boldsymbol{x}) &= \mathbb{E}_{Q_Z}[g_\rho(Z)\mathsf{R}(\boldsymbol{x}, Z)] \\
&= \mathbb{E}_{\rho_Z}[g_\rho(Z)\mathsf{R}(\boldsymbol{x}, Z)] + \mathbb{E}_{Q_Z}[g_\rho(Z)\mathsf{R}(\boldsymbol{x}, Z)] - \mathbb{E}_{\rho_Z}[g_\rho(Z)\mathsf{R}(\boldsymbol{x}, Z)] \\
&= \mathbb{E}_{\rho_Z}[g_\rho(Z)\mathsf{R}(\boldsymbol{x}, Z)] + \int_{\mathcal{Z}} g_\rho(\boldsymbol{z})\mathsf{R}(\boldsymbol{x}, \boldsymbol{z})(\mathrm{d}Q_Z(\boldsymbol{z}) - \mathrm{d}\rho_Z(\boldsymbol{z})).
\end{aligned}$$

Now, unpacking the first term on the right-hand side above, we recognize that for fixed $\boldsymbol{x} \in \mathfrak{X}$, the random variable $\mathsf{R}(\boldsymbol{x}, Z)$ is $\sigma(Z)$-measurable, so via the properties of conditional expectation (Schilling, 2017, Theorem 27.11 (vii)) in $\mathbf{L}^1(\rho_Z)$, we may write

$$\mathbb{E}_{\rho_Z}[g_\rho(Z)\mathsf{R}(\boldsymbol{x}, Z)] = \mathbb{E}_{\rho_Z}\left[\mathbb{E}_{\rho_{Y,Z}}[r(Y)|Z]\,\mathsf{R}(\boldsymbol{x}, Z)\right] = \mathbb{E}_{\rho_Z}\left[\mathbb{E}_{\rho_{Y,Z}}[r(Y)\mathsf{R}(\boldsymbol{x}, Z)|Z]\right].$$

Using the expression above and the tower property of the conditional expectation (Lem. 2), we write

$$\mathbb{E}_{\rho_Z}[g_\rho(Z)\mathsf{R}(\boldsymbol{x}, Z)] = \mathbb{E}_{\rho_Z}\left[\mathbb{E}_{\rho_{Y,Z}}[r(Y)\mathsf{R}(\boldsymbol{x}, Z)|Z]\right] = \mathbb{E}_{\rho_{Y,Z}}[r(Y)\mathsf{R}(\boldsymbol{x}, Z)],$$

completing the proof. $\qquad\square$

Mean Square Contingency. Both singular value decomposition from Prop. 1 and the Radon-Nikodym derivative R from Prop. 2 can be used to calculate a dependence measure between X and Z (Buja, 1990). This dependence measure arises in nonlinear canonical correlation analysis and alternating conditional expectations (Breiman and Friedman, 1985; Bickel et al., 1993).

Definition 8 (Mean Square Contingency). Assume that $\mathbf{M}_{Z|X}$ is Hilbert-Schmidt. Assume that $I = \mathbb{N}$ (Prop. 1), where we may append zeros if I is finite. Recalling that $\sigma_1 = 1$, define the *mean square contingency* $I(X; Z)$ as any of the expressions

$$I(X; Z) := \|\mathbf{M}_{Z|X}\|^2_{\mathrm{HS}(\mathbf{L}^2(Q_Z), \mathbf{L}^2(Q_X))} - 1 = \sum_{i=2}^{\infty} \sigma_i^2.$$

Our definition of the mean square contingency is in fact the square of the quantity that was originally introduced as such by Rényi (1959), which is shown below.

Definition 9. Assume that $Q_{X,Z} \ll Q_X \otimes Q_Z$, so that R exists, and that $\mathsf{R} \in \mathbf{L}^2(Q_X \otimes Q_Z)$. Define the *Rényi mean squared contingency* Rényi (1959, Eq. (13)) as

$$I_{\mathrm{Rényi}}(X; Z) := \|\mathsf{R} - 1\|_{\mathbf{L}^2(Q_X \otimes Q_Z)} = \sqrt{\int_{\mathfrak{X} \times \mathcal{Z}} (\mathsf{R}(\boldsymbol{x}, \boldsymbol{z}) - 1)^2 \, \mathrm{d}(Q_X \otimes Q_Z)(\boldsymbol{x}, \boldsymbol{z})}.$$

If $Q_{X,Z}$ is absolutely continuous with respect to a measure ν on $\mathfrak{X} \times \mathcal{Z}$, with joint density $q_{X,Z}$ and marginal densities (q_X, q_Z), we have that

$$I_{\mathrm{Rényi}}(X; Z) = \sqrt{\int_{\mathfrak{X} \times \mathcal{Z}} (\mathsf{R}(\boldsymbol{x}, \boldsymbol{z}) - 1)^2 \, q_X(\boldsymbol{x}) q_Z(\boldsymbol{z}) \mathrm{d}\nu(\boldsymbol{x}, \boldsymbol{z})}.$$

Written in the form above, $I_{\mathrm{Rényi}}(X; Z)$ may also be called the χ^2-*functional* (Buja, 1990).

We apply $I(X;Z) = I_{\text{Rényi}}(X;Z)^2$ to achieve the sequence of identities following (10). Using the singular decay computations from Lem. 3 with $c = C = 1$, we have that if $\sigma_i = i^{-\gamma}$, then

$$\frac{1}{2\gamma - 1} - 1 \leq I(X;Z) \leq \frac{1}{2\gamma - 1} \iff \frac{1}{2}\frac{I(X;Z) + 2}{I(X;Z) + 1} \leq \gamma \leq \frac{1}{2}\frac{I(X;Z) + 1}{I(X;Z)}. \tag{31}$$

For simplicity, we will use the upper bounds to describe the order of the quantities, that is,

$$I(X;Z) \sim \frac{1}{2\gamma - 1} \iff \gamma \sim \frac{I(X;Z) + 1}{2I(X;Z)}. \tag{32}$$

We employ this relation in the sample complexity calculations in Appx. D.

B.4. Reproducing Kernel Hilbert Spaces

In this section, we review facts about the interplay between reproducing kernel Hilbert spaces (RKHSs), the $\mathbf{L}^2$-spaces defined in Appx. B.1, the Hilbert-Schmidt spaces from Appx. B.2, and Bochner spaces (introduced below). The goal is to provide the necessary background in order to understand the results regarding kernel-based estimation methods that are used in other parts of the manuscript. The analyses in Appx. D rely on being able to decompose some target function (related to the dependence between X and Z) so that it may be estimated in multiple ways. One method involves estimating the Radon-Nikodym derivative R introduced in Prop. 2. The second technique relies on vector-valued regression, with a target function denoted by $F_\star$. Most of the setup below is in service of the vector-valued regression estimation portion.

We maintain the Borel spaces $(\mathfrak{X}, \mathcal{B}(\mathfrak{X}))$ and $(\mathfrak{Z}, \mathcal{B}(\mathfrak{Z}))$ from Appx. B.3, with the topological assumption that $\mathfrak{X}$ and $\mathfrak{Z}$ are second countable, locally compact, and Hausdorff. In addition, $\mathcal{H}$ and $\mathcal{G}$ each denote a separable reproducing kernel Hilbert space (RKHS) containing real-valued functions of $\mathfrak{X}$ and real-valued functions of $\mathfrak{Z}$, respectively. We let $\phi : \mathfrak{X} \to \mathcal{H}$ and $\psi : \mathfrak{Z} \to \mathcal{G}$ be the canonical feature maps and let $k : \mathfrak{X} \times \mathfrak{X} \to \mathbb{R}$ and $l : \mathfrak{Z} \times \mathfrak{Z} \to \mathbb{R}$ be the reproducing kernels for $\mathcal{H}$ and $\mathcal{G}$. The boundedness assumptions $\sup_{\boldsymbol{x}, \boldsymbol{x}' \in \mathfrak{X}} k(\boldsymbol{x}, \boldsymbol{x}') \leq k_{\max} < \infty$ and $\sup_{\boldsymbol{z}, \boldsymbol{z}' \in \mathfrak{Z}} l(\boldsymbol{z}, \boldsymbol{z}') \leq l_{\max} < \infty$ are maintained throughout the paper.

Bochner Space. We will adopt the equivalence class notation first introduced in Appx. B.1, with respect to probability measures Q_X and Q_Z. That is, for any two real-valued measurable functions $f : \mathfrak{X} \to \mathbb{R}$ and $h : \mathfrak{X} \to \mathbb{R}$, we say that $f \sim_X h$ if

$$Q_X (\{\boldsymbol{x} \in \mathfrak{X} : f(\boldsymbol{x}) \neq h(\boldsymbol{x})\}) = 0.$$

The notation $[f]_X$ denotes an equivalence class with respect to the equivalence relation $\sim_X$, with representative f. We say that $[f]_X \in \mathbf{L}^2(Q_X)$ if

$$\int_{\mathfrak{X}} h^2(\boldsymbol{x}) \mathrm{d}Q_X(\boldsymbol{x}) < \infty \text{ for some, or equivalently all } h \in [f]_X.$$

We define $\sim_Z$, $[\cdot]_Z$, and $\mathbf{L}^2(Q_Z)$ similarly. We introduce a similar construction to $\mathbf{L}^2(Q_X)$ for *vector-valued* functions, i.e., those whose outputs lie in a Hilbert space. For measurable functions $F : \mathfrak{X} \to \mathcal{G}$ and $H : \mathfrak{X} \to \mathcal{G}$, we will define the equivalence relation $F \sim_X H$ via $Q_X (\{\boldsymbol{x} \in \mathfrak{X} : F(\boldsymbol{x}) \neq H(\boldsymbol{x})\}) = 0$, and corresponding equivalence classes will be denoted $[F]_X$. We define the *Bochner space* $\mathbf{L}^2(Q_X; \mathcal{G})$ via $[F]_X \in \mathbf{L}^2(Q_X; \mathcal{G})$ if

$$\int_{\mathfrak{X}} \|H(\boldsymbol{x})\|_{\mathcal{G}}^2 \mathrm{d}Q_X(\boldsymbol{x}) < \infty \text{ for some, or equivalently all } H \in [F]_X.$$

Analogous to $\mathbf{L}^2(Q_X)$, this is a set of equivalence classes of vector-valued functions. Recall from Appx. B.2 and Appx. B.3 that we use $\mathrm{HS}(\mathcal{U}, \mathcal{V})$ to denote the space of Hilbert-Schmidt operators mapping from a Hilbert space $\mathcal{U}$ to another Hilbert space $\mathcal{V}$. The following result allows us to relate elements of the Bochner space $\mathbf{L}^2(Q_X; \mathcal{G})$ to elements of a space of Hilbert-Schmidt operators $\mathrm{HS}(\mathbf{L}^2(Q_X), \mathcal{G})$. For $[f]_X \in \mathbf{L}^2(Q_X)$ and $g \in \mathcal{G}$, the notation $f(\cdot)g$ refers to the function mapping $\boldsymbol{x} \in \mathfrak{X}$ to $f(\boldsymbol{x})g \in \mathcal{G}$.

Theorem 6. *(Aubin, 2000, Theorem 12.6.1) There exists a function $\Phi : \mathrm{HS}(\mathbf{L}^2(Q_X), \mathcal{G}) \to \mathbf{L}^2(Q_X; \mathcal{G})$ that is a bijective linear transformation satisfying*

$$\|\mathbf{C}\|_{\mathrm{HS}(\mathbf{L}^2(Q_X), \mathcal{G})} = \|\Phi(\mathbf{C})\|_{\mathbf{L}^2(Q_X; \mathcal{G})} \text{ for all } \mathbf{C} \in \mathrm{HS}(\mathbf{L}^2(Q_X), \mathcal{G}),$$

and for every $[f]_X \in \mathbf{L}^2(Q_X)$ and $g \in \mathcal{G}$, associates

$$\Phi(g \otimes [f]_X) = [f(\cdot)g]_X \iff g \otimes [f]_X = \Phi^{-1}([f(\cdot)g]_X) \tag{33}$$

for the rank-one operator $g \otimes [f]_X \in \mathrm{HS}(\mathbf{L}^2(Q_X), \mathcal{G})$.

Based on the definition of the Hilbert-Schmidt norm in (25) (Appx. B.2, Thm. 6 will make computation of $\mathbf{L}^2(Q_X; \mathcal{G})$-norms more convenient by relating them to $\mathrm{HS}(\mathbf{L}^2(Q_X), \mathcal{G})$-norms. The following technical lemma can be used to simplify computations regarding the inverse of this isomorphism.

Lemma 7. *Let $(g_j)_{j \in J}$ be any countable orthonormal basis of $\mathcal{G}$, and let $\mathbf{C} = \Phi^{-1}([F]_X)$ for some $F : \mathcal{X} \to \mathcal{G}$ such that $[F]_X \in \mathbf{L}^2(Q_X; \mathcal{G})$. Define the functions $(f_j)_{j \in J}$ via $f_j(\boldsymbol{x}) := \langle g_j, F(\boldsymbol{x}) \rangle_{\mathcal{G}}$. Then, $[f_j]_X \in \mathbf{L}^2(Q_X)$ for all $j \in J$, and we have the identity*

$$\mathbf{C} = \sum_{j \in J} g_j \otimes [f_j]_X,$$

where the convergence is interpreted in terms of $\mathrm{HS}(\mathbf{L}^2(Q_X), \mathcal{G})$.

Proof. First, consider the case in which we can write the equivalence class of F in $\mathbf{L}^2(Q_X, \mathcal{G})$ in the form

$$[F]_X = \sum_{j \in J} [f_j(\cdot)g_j]_X, \tag{34}$$

for some sequence of functions $f_1, f_2, \ldots \in \mathbf{L}^2(Q_X)$. Then, because Φ^{-1} is a linear isometry, it is a bounded (hence continuous) operator with respect to the norm on $\mathbf{L}^2(Q_X, \mathcal{G})$. This implies via continuity

$$\mathbf{C} = \Phi^{-1}([F]_X) = \Phi^{-1}\left(\sum_{j \in J} [f_j(\cdot)g_j]_X\right) = \sum_{j \in J} \Phi^{-1}([f_j(\cdot)g_j]_X) = \sum_{j \in J} g_j \otimes [f_j]_X,$$

where the last step follows because Φ^{-1} satisfies the relation (33). To achieve the identity (34), we fix any $\boldsymbol{x} \in \mathcal{X}$, we expand $F(\boldsymbol{x}) \in \mathcal{G}$ onto the basis $(g_j)_{j \in J}$ to write

$$F(\boldsymbol{x}) = \sum_{j \in J} \underbrace{\langle g_j, F(\boldsymbol{x}) \rangle_{\mathcal{G}}}_{f_j(\boldsymbol{x})} g_j.$$

To pass this pointwise equality to (34), consider any sequence of $\mathcal{G}$-valued functions $(H_j)_{j \in J}$ such that $H_j(\boldsymbol{x}) = f_j(\boldsymbol{x})g_j$ for all $\boldsymbol{x} \in \mathcal{X}_j \subseteq \mathcal{X}$, where $Q_X(\mathcal{X}_j) = 1$. Similarly, consider $H_0 : \mathcal{X} \to \mathcal{G}$ such that $H_0(\boldsymbol{x}) = F(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}_0$ with $Q_X(\mathcal{X}_0) = 1$. Thus, we have that

$$H_0(\boldsymbol{x}) = \sum_{j \in J} H_j(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \in \mathcal{X}_0 \cap \left(\bigcap_{j \in J} \mathcal{X}_j\right),$$

and because J is countable, this implies that $H_0(\boldsymbol{x}) = \sum_{j \in J} H_j(\boldsymbol{x})$ for Q_X-almost all $\boldsymbol{x} \in \mathcal{X}$, granting (34). It remains to be shown that $[f_j]_X \in \mathbf{L}^2(Q_X)$. This follows by the Bochner-square integrability of F, as

$$\int_{\mathcal{X}} f_j^2(\boldsymbol{x}) \mathrm{d}Q_X(\boldsymbol{x}) = \int_{\mathcal{X}} \langle g_j, F(\boldsymbol{x}) \rangle_{\mathcal{G}}^2 \mathrm{d}Q_X(\boldsymbol{x}) \leq \|g_j\|_{\mathcal{G}}^2 \int_{\mathcal{X}} \|F(\boldsymbol{x})\|_{\mathcal{G}}^2 \mathrm{d}Q_X(\boldsymbol{x}) < \infty,$$

which completes the proof. $\square$

In the sequel, we will define a statistical learning problem in which the target function $F_\star$ is an element of $\mathbf{L}^2(Q_X; \mathcal{G})$.

For the kernel-based estimation approach, the estimation function $\widehat{F} \equiv \widehat{F}_\lambda$ (where λ is a to-be-specified regularization parameter) will live in a particular vector-valued RKHS that will be isometrically isomorphic to $\mathrm{HS}(\mathcal{H}, \mathcal{G})$. Using Thm. 6, $F_\star$ will be associated to en element $\mathbf{C}_\star \in \mathrm{HS}(\mathbf{L}^2(Q_X), \mathcal{G})$ via an isometric isomorphism. We introduce the concept of embeddings and interpolation spaces to describe exactly where $\mathbf{C}_\star$ lies in between $\mathrm{HS}(\mathcal{H}, \mathcal{G})$ and $\mathrm{HS}(\mathbf{L}^2(Q_X), \mathcal{G})$ (via a *source condition*).

Embedding Operator. See Appx. B.2 for a review of the terminology surrounding compact operators. Consider the *embedding operator* $\mathbf{I}_X : \mathcal{H} \to \mathbf{L}^2(Q_X)$, which identifies a function $h \in \mathcal{H}$ with its equivalence class $[h]_X \in \mathbf{L}^2(Q_X)$. Under the bounded kernel assumption, we have that $\mathbf{I}_X$ is compact, and moreover Hilbert-Schmidt, with norm bounded as $\|\mathbf{I}_X\|_{\mathrm{HS}(\mathcal{G}, \mathbf{L}^2(Q_X))} \leq \sqrt{k_{\max}}$ (Steinwart and Scovel, 2012, Lemma 2.3). We denote its adjoint by $\mathbf{S}_X := \mathbf{I}_X^*$, and finally, construct the self-adjoint, trace class operator

$$\mathbf{T}_X := \mathbf{I}_X \mathbf{S}_X : \mathbf{L}^2(Q_X) \to \mathbf{L}^2(Q_X). \tag{35}$$

Applying the eigendecomposition Thm. 4, there exists an orthonormal basis of $\mathrm{cl}(\mathrm{range}(\mathbf{I}_X)) \subseteq \mathbf{L}^2(Q_X)$, denoted $([e_{X,i}]_X)_{i \in I}$, and a sequence of positive, non-increasing eigenvalues $(\mu_{X,i})_{i \in I}$ such that

$$\mathbf{T}_X = \sum_{i \in I} \mu_{X,i} \langle [e_{X,i}]_X, \cdot \rangle_{\mathbf{L}^2(Q_X)} [e_{X,i}]_X. \tag{36}$$

Note that we have only used the index set I from Thm. 4, as opposed to the larger set J for which we can define an ONB for the entirety of $\mathbf{L}^2(Q_X)$, not only $\mathrm{cl}(\mathrm{range}(\mathbf{I}_X))$. Analogous to $\mathbf{T}_X$, we can also define the uncentered covariance operator

$$\mathbf{C}_X = \mathbf{S}_X \mathbf{I}_X : \mathcal{H} \to \mathcal{H}.$$

Similar to (36), $\mathbf{C}_X$ enjoys an eigendecomposition

$$\mathbf{C}_X = \sum_{i \in I} \mu_{Z,i} \langle \cdot, \mu_{X,i}^{1/2} e_{X,i} \rangle_{\mathcal{G}} \mu_{X,i}^{1/2} e_{X,i}. \tag{37}$$

The equation above implicitly contains another fact, which is that the equivalence classes in (36) all contain representatives that are in $\mathcal{H}$. This defines the collection $\{e_{X,i}\}_{i \in I}$, which forms an ONB of $\mathrm{null}(\mathbf{I}_X)^\perp \subseteq \mathcal{H}$. Combining (36) and (37), the embedding can be described using a singular value decomposition

$$\mathbf{I}_X = \sum_{i \in I} \mu_{X,i}^{1/2} \left([e_{X,i}]_Z \otimes (\mu_{X,i}^{1/2} e_{X,i}) \right). \tag{38}$$

Lastly, we define $(\mathbf{I}_Z, \mathbf{S}_Z, \mathbf{T}_Z, \mathbf{C}_Z)$ as the analogous operators for $\mathbf{L}^2(Q_Z)$ and $\mathcal{G}$.

Interpolation Spaces and the Inclusion Map. For any $\alpha \geq 0$, we define the operator

$$\mathbf{T}_X^{\alpha/2} = \sum_{i \in I} \mu_{X,i}^{\alpha/2} \langle [e_{X,i}]_X, \cdot \rangle_{\mathbf{L}^2(Q_X)} [e_{X,i}]_X, \tag{39}$$

$$\mathrm{dom}(\mathbf{T}_X^{\alpha/2}) = \left\{ [f]_X \in \mathbf{L}^2(Q_X) : \sum_{i \in I} \mu_{X,i}^{\alpha/2} \langle [f]_X, [e_{X,i}]_X \rangle_{\mathbf{L}^2(Q_X)} < \infty \right\},$$

which is considered to be well-defined when $\mathrm{dom}(\mathbf{T}_X^{\alpha/2}) \neq \varnothing$. Then, we define the *$\alpha$-interpolation space* $[\mathcal{H}]^\alpha$ via

$$[\mathcal{H}]^\alpha = \left\{ \sum_{i \in I} a_i \mu_{X,i}^{\alpha/2} [e_{X,i}]_X : (a_i)_{i \in I} \in \ell_2(I) \right\} \subseteq \mathbf{L}^2(Q_X),$$

where $(a_i)_{i \in I} \in \ell_2(I)$ indicates that $\sum_{i \in I} a_i^2 < +\infty$. When $\alpha = 0$, we recover $[\mathcal{H}]^0 = \mathrm{cl}(\mathrm{range}(\mathbf{I}_X))$, whereas for $\alpha = 1$, $[\mathcal{H}]^1$ is isometrically isomorphic to $\mathrm{null}(\mathbf{I}_X)^\perp \subseteq \mathcal{H}$ (Fischer and Steinwart, 2020). Thus, for $\alpha \in (0, 1)$, we interpret $[\mathcal{H}]^\alpha$ as an "interpolation" between the well-behaved functions in the RKHS $\mathcal{H}$ and the elements of $\mathbf{L}^2(Q_X)$. Associated to each

$[\mathcal{H}]^\alpha$ is the *inclusion map* $\mathbf{I}_X^{\alpha,\infty}$, which simply views an element $[h]_X \in [\mathcal{H}]^\alpha$ as an element of $\mathbf{L}^\infty(Q_X)$ (this requires the boundedness of the kernel). Here, $\mathbf{L}^\infty(Q_X)$ denotes equivalence classes of real-valued functions on $\mathcal{X}$ that have a finite essential supremum under Q_X. We write $\mathbf{I}_X^{\alpha,\infty} : [\mathcal{H}]^\alpha \hookrightarrow \mathbf{L}^\infty(Q_X)$ when the inclusion map is continuous (see Asm. 10).

We use the standard generalization of these notions onto spaces of vector-valued functions (Li et al., 2024; Meunier et al., 2024): for any $\beta \geq 0$, we define the β-*interpolation norm* for $\mathbf{C} \in \mathrm{HS}(\mathbf{L}^2(Q_X), \mathcal{G})$ via the formula

$$\|\mathbf{C}\|_\beta := \|\mathbf{C}\mathbf{T}_X^{-\beta/2}\|_{\mathrm{HS}(\mathbf{L}^2(Q_X), \mathcal{G})} \in [0, +\infty]. \tag{40}$$

This norm, when finite, will be used to define the source condition of the target function $F_\star$ alluded to in Sec. 3, as we may compute $\|\mathbf{C}_\star\|_\beta$ for $\mathbf{C}_\star := \Phi^{-1}([F_\star]_X)$ (see Thm. 6). While we phrase the condition in terms of the constant β above in order to relate it to the kernel methods and inverse problem literature below, we will use the constructions of Appx. B.3 to phrase the finiteness of (40) for $\mathbf{C}_\star$ in terms of the mean square contingency (Definition 8) in Appx. D.

Vector-Valued Spectral Regularization Learning. We may now describe estimation techniques for an $\mathbf{L}^2(Q_X; \mathcal{G})$-valued target function that fall into the category of vector-valued spectral regularization learning. We give only a brief overview in order to state the statistical convergence guarantees; see Meunier et al. (2024) for a detailed description, including computational properties of the estimator. As we prove in Lem. 8, there exists a function $F_\star : \mathcal{X} \to \mathcal{G}$ such that $[F_\star]_X \subseteq \mathbf{L}^2(Q_X; \mathcal{G})$ and for every $g \in \mathcal{G}$,

$$\mathbb{E}_{Q_{X,Z}}[g(Z)|X](\boldsymbol{x}) = \langle g, F_\star(\boldsymbol{x})\rangle_\mathcal{G}. \tag{41}$$

For each $\boldsymbol{x} \in \mathcal{X}$, we also refer to $F_\star(\boldsymbol{x})$ as the conditional mean embedding of Z given $X = \boldsymbol{x}$. Note that for a fixed $g \in \mathcal{G}$, we do not assume that $\boldsymbol{x} \mapsto \langle g, F_\star(\boldsymbol{x})\rangle$ is an element of an RKHS $\mathcal{H}$. This avoids some of the technical challenges raised, for instance, in Klebanov et al. (2020; 2021), where this requirement places strong implicit restrictions on the chosen kernel and RKHS. Instead, the mis-specified case is handled using vector-valued interpolation spaces.

Next, using Thm. 6, we associate to $F_\star$ the element $\mathbf{C}_\star = \Phi^{-1}(F_\star) \in \mathrm{HS}(\mathbf{L}^2(Q_X), \mathcal{G})$. Given independent and identically distributed pre-training data $(X_1, Z_1), \ldots, (X_N, Z_N)$ drawn from $Q_{X,Z}$, define the empirical (uncentered) auto-covariance and cross-covariance operator

$$\widehat{\mathbf{C}}_{XX} = \frac{1}{N}\sum_{i=1}^N \phi(X_i) \otimes \phi(X_i) \text{ and } \widehat{\mathbf{C}}_{ZX} = \frac{1}{N}\sum_{i=1}^N \psi(Z_i) \otimes \phi(X_i).$$

Let $f_\lambda : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ denote the spectral cutoff function

$$f_\lambda(x) = \begin{cases} x^{-1} & \text{if } x \geq \lambda \\ 0 & \text{otherwise} \end{cases}. \tag{42}$$

We can interpret $f_\lambda(x)$ as a regularized inverse that behaves in a reasonable manner near $x = 0$. A similar function corresponding to the more familiar Tikhonov regularization is $f_\lambda(x) = (x + \lambda)^{-1}$. While other options for f_λ (i.e. filter functions) exist owing to the tools of regularization theory (Bauer et al., 2007), the spectral cutoff function will be sufficient for our purposes, as it allows for the simplest statement of the upcoming results. For a self-adjoint positive semidefinite operator $\mathbf{C}$, we define $f_\lambda(\mathbf{C})$ as replacing each eigenvalue $\mu_i \geq 0$ of $\mathbf{C}$ with $f_\lambda(\mu_i)$ in the eigendecomposition (see Thm. 4). For regularization parameter $\lambda > 0$, we define the estimator

$$\widehat{F}_\lambda(\cdot) := \widehat{\mathbf{C}}_\lambda \phi(\cdot) \text{ for } \widehat{\mathbf{C}}_\lambda := \widehat{\mathbf{C}}_{ZX} f_\lambda(\widehat{\mathbf{C}}_{XX}) : \mathcal{H} \to \mathcal{G}. \tag{43}$$

Now, consider the following assumptions, which include the source condition.

Assumption 10. (Meunier et al., 2024, Assumptions (SRC), (MOM), (EVD), (EMB))

1. There exist positive constants $\beta > 0$ and $B > 0$ such that $\|F_\star\|_\beta := \|\mathbf{C}_\star\|_\beta \leq B$.

2. For positive constants $\sigma^2, c > 0$ the Bernstein moment condition

$$\mathbb{E}_{Q_{X,Z}}\left[\|\psi(Z) - F_\star(X)\|_\mathcal{G}^2 | X\right](\boldsymbol{x}) \leq \frac{1}{2}q!\sigma^2 c^{q-2}$$

   is satisfied for Q_X-almost all $\boldsymbol{x} \in \mathcal{X}$ and all $q \geq 2$.

3. There exist constants $D > 0$ and $p < 1$ such that

$$\mu_{X,i} \leq Di^{-1/p}.$$

4. For $\alpha \in [p, 1]$, the inclusion map $\mathbf{I}_X^{\alpha,\infty} : [\mathcal{H}]^\alpha \hookrightarrow \mathbf{L}^\infty(Q_X)$ is bounded, with operator norm $\|I_X^{\alpha,\infty}\|_{\mathrm{op}} \leq A$.

Note that the first assumption is always satisfied for $\alpha = 1$, due to boundedness of the kernel (as the $[\mathcal{H}]^1$ norm can be associated to the RKHS norm of an element of $\mathcal{H}$). We pay particular attention to the constant β which defines the aforementioned source condition. In Appx. D.1, we translate this condition into one regarding the dependence between X and Z, using the tools from Appx. B.3. We refer to the case when $\beta \geq 1$ as the *well-specified* case. We also employ one additional assumption to state the result.

Assumption 11 (Sub-Gaussian Tail). There exists a positive constant $\tau > 0$ such that the following holds:

$$\mathbb{P}_{Q_X}\left[\|F_\star(X)\|_\mathcal{G} > t\right] \leq 2e^{-\frac{t^2}{2\nu^2}}.$$

Asm. 11 is only used to replace a statement of the form "for $N \geq 1$ sufficiently large" from Meunier et al. (2024, Theorem 4) with a quantitative condition on N. It is used to control the probability that $\|F_\star(X_i)\|_\mathcal{G} > t$ for any $i = 1, \ldots, N$ for the choices of t specified in the proof of Meunier et al. (2024, Theorem 8).

Theorem 7. *(Meunier et al., 2024, Theorem 4) Consider a failure probability $\delta \in (0, 1]$, the estimate $\widehat{F}_\lambda$ defined in (43), and the target function $F_\star$ defined in (41). Under Asm. 10 and Asm. 11, there exists a constant $C > 0$ (independent of N and δ) such that the following statements hold.*

- **Case 1:** $\beta + p > \alpha$. *If $N^{\left(\frac{1}{2}\left(1 + \frac{p-\alpha}{p+\beta}\right) + \frac{\alpha-\beta}{2\alpha}\right)} \geq 2\nu^2 \operatorname{plog}(N/\delta)$ and $\lambda = \Theta(N^{-\frac{1}{\beta+p}})$, then*

$$\|[\widehat{F}_\lambda]_X - F_\star\|_{\mathbf{L}^2(Q_X;\mathcal{G})}^2 \leq C \operatorname{plog}(1/\delta) N^{-\frac{\beta}{\beta+p}}$$

   with probability at least $1 - \delta/4$.

- **Case 2:** $\beta + p \leq \alpha$. *If $N^{\frac{\alpha-\beta}{\alpha}} \geq 2\nu^2 \operatorname{plog}(N/\delta)$ and $\lambda = \Theta((N/\operatorname{plog}(N))^{-\frac{1}{\alpha}})$, then*

$$\|[\widehat{F}_\lambda]_X - F_\star\|_{\mathbf{L}^2(Q_X;\mathcal{G})}^2 \leq C \operatorname{plog}(1/\delta)(N/\log^2(N))^{-\frac{\beta}{\alpha}}$$

   with probability at least $1 - \delta/4$.

This result is applied in Appx. D.1 and provides an example of the "conditional mean" approach outlined in Sec. 2 and Sec. 3. Regarding the setting of λ in Thm. 7, the argument follows the typical recipe of defining an element $F_\lambda \in \mathbf{L}^2(Q_X; \mathcal{G})$ which represents the population version of the regularized predictor. Let $\|\cdot\|_\gamma$ denote the γ-interpolation norm for $\gamma \in [0, 1]$, which is equal to $\|\cdot\|_{\mathbf{L}^2(Q_X;\mathcal{G})}$ when $\gamma = 0$. Then, the approximation error $\|[F_\lambda]_X - F_\star\|_\gamma^2$ decays according to $\lambda^{\beta-\gamma}$ when using the spectral cutoff regularizer, which reflects the classical analyses of Smale and Zhou (2007). In the well-specified case, the estimation error, $\|[\widehat{F}_\lambda]_X - [F_\lambda]_X\|_\gamma$ decomposes into multiple terms which include irreducible noise terms of order $N^{-1}\lambda^{-\alpha/2}$ and additional approximation terms of order $N^{-1/2}\lambda^{(\beta-\alpha)/2}$. By using $\lambda = \Theta(N^{-\frac{1}{\beta+p}})$ and the condition $\beta + p > \alpha$ from Case 1, the irreducible noise error converges at rate $N^{-1/2}$ whereas the approximation term converges at rate $N^{-\beta/2(\beta+p)}$. Note that these rates will be squared in Thm. 7. The argument for Case 2 follows similarly.

Radon-Nikodym Derivative Estimation. To set the stage for this technique, we describe a function class in which $\widehat{\mathsf{R}} : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ will live. Let $\mathcal{S}$ denote a separable reproducing kernel Hilbert space (RKHS) of real-valued functions

on $\mathcal{X} \times \mathcal{Z}$, with canonical feature map $\varphi : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ and reproducing kernel $\kappa : (\mathcal{X} \times \mathcal{Z}) \times (\mathcal{X} \times \mathcal{Z}) \to \mathbb{R}$. As before, we first assume boundedness of the kernel, i.e., $\sup \{\kappa(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{x}', \boldsymbol{z}') : (\boldsymbol{x}, \boldsymbol{z}), (\boldsymbol{x}', \boldsymbol{z}') \in \mathcal{X} \times \mathcal{Z}\} \leq \kappa_{\max}$.

Let us describe the estimation procedure, which relies on a similar spectral regularization technique as the one described for vector-valued regression. Because the Radon-Nikodym derivative being estimated is $\frac{\mathrm{d}Q_{X,Z}}{\mathrm{d}(Q_X \otimes Q_Z)}$, we consider having samples from both distributions available. In particular, we observe N_p paired examples $(X_1, Z_1), \ldots, (X_{N_\mathrm{p}}, Z_{N_\mathrm{p}}) \sim Q_{X,Z}$ and N_u unpaired examples $(X_1', Z_1'), \ldots, (X_{N_\mathrm{u}}', Z_{N_\mathrm{u}}') \sim Q_X \otimes Q_Z$. Define the uncentered covariance operators

$$\widehat{\mathbf{C}}_\mathrm{p} = \frac{1}{N_\mathrm{p}} \sum_{i=1}^{N_\mathrm{p}} \varphi(X_i, Z_i) \otimes \varphi(X_i, Z_i), \quad \widehat{\mathbf{C}}_\mathrm{u} = \frac{1}{N_\mathrm{u}} \sum_{i=1}^{N_\mathrm{u}} \varphi(X_i', Z_i') \otimes \varphi(X_i', Z_i'), \tag{44}$$

representing the paired and unpaired examples, respectively. Then, using the spectral cutoff function f_λ (see (42)), we define the estimate

$$\widehat{\mathsf{R}} \equiv \widehat{\mathsf{R}}_\lambda = f_\lambda(\widehat{\mathbf{C}}_\mathrm{u}) \widehat{\mathbf{C}}_\mathrm{p} \mathbf{1}, \tag{45}$$

where $\mathbf{1}(\boldsymbol{x}, \boldsymbol{z}) = 1$ for all $(\boldsymbol{x}, \boldsymbol{z}) \in \mathcal{X} \times \mathcal{Z}$. Because f_λ can be viewed as a regularized inverse, $\widehat{\mathsf{R}}$ can readily be interpreted as the "ratio" of the covariance operator of the paired sample over that of the unpaired sample.

To state the assumptions for the analysis, we require an analogous operator to $\mathbf{I}_X$ and $\mathbf{I}_Z$ introduced earlier in this section. We then define the *embedding operator* $\mathbf{I}_{X,Z} : \mathcal{S} \to \mathbf{L}^2(Q_X \otimes Q_Z)$, which takes an element $\mathsf{S} \in \mathcal{S}$ and indexes its equivalence class in $\mathbf{L}^2(Q_X \otimes Q_Z)$. We will not need to define an explicit notation for the equivalence class for this discussion, but will do so in Appx. D.2. Due to Steinwart and Scovel (2012, Lemma 2.3), the bounded kernel assumption implies that $\mathbf{I}_{X,Z}$ is Hilbert-Schmidt with norm bounded as $\|\mathbf{I}_{X,Z}\|_{\mathrm{HS}(\mathcal{S}, \mathbf{L}^2(Q_X \otimes Q_Z))} < \sqrt{\kappa_{\max}}$.

Recall the powers of operators introduced in (39). We will use a similar construction for this technique as well. Define the (compact) adjoint operator $\mathbf{S}_{X,Z} = \mathbf{I}_{X,Z}^* : \mathbf{L}^2(Q_X \otimes Q_Z) \to \mathcal{S}$. Via Thm. 4, let $(\mu_i)_{i \in I}$ denote the non-zero eigenvalues of the compact, trace class operator $\mathbf{S}_{X,Z} \mathbf{I}_{X,Z}$, where we consider $I = \mathbb{N}$ for simplicity. Let the degrees of freedom function be defined as

$$\mathrm{df}(\lambda) := \sum_{i=1}^{\infty} \frac{\mu_i}{\mu_i + \lambda}.$$

Consider the following assumption.

Assumption 12. (Nguyen et al., 2024, Eq. (9) and Remark 13) There exists an absolute constant C_df and a constant $\alpha > 1$ such that $\mathrm{df}(\lambda) \leq C_\mathrm{df} \lambda^{-1/\alpha}$. There exists a $\beta \geq 1$, along with an element $\mathsf{S}_{Q_{X,Z}} \in \mathrm{null}(\mathbf{I}_{X,Z})^\perp$, such that

$$\mathsf{R} = (\mathbf{S}_{X,Z} \mathbf{I}_{X,Z})^\beta \mathsf{S}_{Q_{X,Z}}.$$

The upper bound on $\mathrm{df}(\lambda)$ reflects a polynomial eigendecay of order $\mu_i \sim i^{-\alpha}$ (see Bach (2024, Section 7.6.6)). Asm. 12 is more specific than the one stated in the referenced work, in that we use the specific index function $x \mapsto x^\beta$, growing at least linearly. Furthermore, their result may achieve faster convergence rates than the one stated in Cor. 2 using an additional source condition on the feature map φ. However, our intention is not necessarily to provide convergence rates that are optimal in a particular parameter regime, but ones that are informative with regard to the dependence structure of $Q_{X,Z}$. To this end, we do not incorporate the additional condition.

To state the result, we define $\lambda_\star$ as the solution of

$$\frac{\mathrm{df}(\lambda)}{\lambda} = N_\mathrm{u},$$

which is guaranteed to exist as $\frac{\mathrm{df}(\lambda)}{\lambda}$ is decreasing from $+\infty$ to 0 on the interval $(0, +\infty)$. Observe the following.

Theorem 8. *(Nguyen et al., 2024, Proposition 10 and Lemma 11) Consider a failure probability $\delta \in (0, 1]$ and constant β*

from Asm. 12. Consider the estimate $\widehat{\mathsf{R}} \equiv \widehat{\mathsf{R}}_\lambda$ defined in (45) and the target function R defined in (2). Finally, define

$$K_{\max} := 1 + (4\kappa_{\max}^2 + \kappa_{\max})^2.$$

There exists a constant $C > 0$ (independent of N and δ) such that for all $\lambda \in [\lambda_\star, \kappa_{\max}]$,

$$\|\widehat{\mathsf{R}} - \mathsf{R}\|_{\mathcal{S}} \le C\,\mathrm{plog}(1/\delta) \left[K_{\max}\lambda^\beta + \frac{K_{\max}^{1/2}}{\lambda} \left(N_{\mathrm{p}}^{-1/2} + N_{\mathrm{u}}^{-1/2} \right) \right] \tag{46}$$

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

By optimizing the bound appearing in (46) in λ, we get that

$$\lambda \equiv \lambda_{N_{\mathrm{p}}, N_{\mathrm{u}}} = \left(\frac{N_{\mathrm{p}}^{-1/2} + N_{\mathrm{u}}^{-1/2}}{K_{\max}^{1/2}} \right)^{\frac{1}{\beta+1}}. \tag{47}$$

If the expression from (47) falls within $[\lambda_\star, \kappa_{\max}]$, this yields the upper bound

$$\|\widehat{\mathsf{R}} - \mathsf{R}\|_{\mathcal{S}} \le C\,\mathrm{plog}(1/\delta) \left[K_{\max}^{\frac{\beta+2}{2(\beta+1)}} \left(N_{\mathrm{p}}^{-1/2} + N_{\mathrm{u}}^{-1/2} \right)^{\frac{\beta}{\beta+1}} \right]. \tag{48}$$

The condition $\lambda_{N_{\mathrm{p}}, N_{\mathrm{u}}} \le \kappa_{\max}$ can be satisfied by taking $(N_{\mathrm{p}}, N_{\mathrm{u}})$ sufficiently large. For the condition that $\lambda_{N_{\mathrm{p}}, N_{\mathrm{u}}} \ge \lambda_\star$, we find an upper bound on $\lambda_\star$ by first deriving an upper bound on $\mathrm{df}(\lambda)/\lambda$, and then solving the resulting equation in λ. By Asm. 12, we have that

$$\frac{\mathrm{df}(\lambda)}{\lambda} \le C_{\mathrm{df}} \lambda^{-(\alpha+1)/\alpha} \implies \lambda_\star \le \left(\frac{C_{\mathrm{df}}}{N_{\mathrm{u}}} \right)^{\frac{\alpha}{\alpha+1}}. \tag{49}$$

Viewing the dependence of (47) on N_{u}, we see that if $\beta \ge (1 - \alpha)/(2\alpha)$, then there exists N_{u} large enough such that (47) is greater than the right-hand side of (49). This is always satisfied, as $\alpha > 1$ is required for $\mathbf{S}_{X,Z}\mathbf{I}_{X,Z}$ to be trace class. Thus, we have the following convergence rate.

Corollary 2. *Adopt the setting of Thm. 8. Let N_{u} be large enough such that (47) is upper bounded by $\kappa_{\max}$ and lower bounded by the right-hand side of (49). Then, for the choice (47), it holds that*

$$\|\widehat{\mathsf{R}} - \mathsf{R}\|_{\mathcal{S}}^2 \le C\,\mathrm{plog}(1/\delta) \left[K_{\max}^{\frac{\beta+2}{\beta+1}} \left(N_{\mathrm{p}}^{-1/2} + N_{\mathrm{u}}^{-1/2} \right)^{\frac{2\beta}{\beta+1}} \right]. \tag{50}$$

This result is applied in Appx. D.2 and provides an example of the "information density" approach outlined in Sec. 2 and Sec. 3. In the sequel, we will simply assume that $N_{\mathrm{p}} = N_{\mathrm{u}} = N/2$ to simplify the statement of the result. Finally, note that the source condition Asm. 12 does not have any implications for the mis-specified case ($\mathsf{R} \notin \mathcal{S}$). This aspect of Radon-Nikodym estimation methodology is still an active area of research in statistical learning.

C. Prompt Bias and Residual Dependence

This appendix is dedicated to the proof of Thm. 1, which controls the population quantity $\|\eta_\star - \eta_\rho\|_{\mathbf{L}^2(P_X)}^2$. The result will follows from Thm. 9, which is a more mathematically precise version of Thm. 1 from the main text.

We recall the problem setting of Sec. 3. We consider the three central probability measures $P_{X,Y}$ (evaluation distribution), $Q_{X,Z}$ (pre-training distribution), and $\rho_{Y,Z}$ (prompt distribution). We notice that $\eta_\star$ (from (3)) depends on $P_{X,Y}$, while η_ρ (from (4)) depends on the pair $(Q_{X,Z}, \rho_{Y,Z})$. Neither component of this term depends on a joint probability over $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. Thus, in order to relate them on common ground, we consider a joint probability measure $P_{X,Y,Z}$, which satisfies certain constraints that make it compatible with the distributions that have observable data. We call this the *latent caption model*.

To proceed, we will need to make several mild regularity conditions on $P_{X,Y,Z}$. We use the notion of regular conditional distribution, or r.c.d. (Definition 4), introduced in Appx. B.1. We use more explicit notation in this section (e.g. $Z = z$) as

compared to Sec. 3 to emphasize the random variable being conditioned on. The assumption below provides a more formal description of Asm. 1 from Sec. 3.

Assumption 13. The joint probability $P_{X,Y,Z}$ on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ satisfies the following constraints.

- **Agrees jointly with the evaluation distribution:** For all measurable sets $A \subseteq \mathcal{X} \times \mathcal{Y}$, we have that $P_{X,Y,Z}(A \times \mathcal{Z}) = P_{X,Y}(A)$ (i.e. $P_{X,Y,Z}$ agrees with the given marginal $P_{X,Y}$).

- **Agrees conditionally with the pre-training distribution:** There exists a measurable set $\mathcal{X}_1 \subseteq \mathcal{X}$ with $P_X(\mathcal{X}_1) = 1$ such that the regular conditional distributions $Q_{Z|X=\boldsymbol{x}}$ and $P_{Z|X=\boldsymbol{x}}$ on $\mathcal{Z}$ exist. Furthermore, these satisfy $Q_{Z|X=\boldsymbol{x}} = P_{Z|X=\boldsymbol{x}}$ for all $\boldsymbol{x} \in \mathcal{X}_1$.

- **Regularity of conditional distributions:** There exists a measurable set $\mathcal{Z}_1 \subseteq \mathcal{Z}$ with $P_Z(\mathcal{Z}_1) = 1$ such that the regular conditional distributions $P_{X,Y|Z=\boldsymbol{z}}$ on $\mathcal{X} \times \mathcal{Y}$ exists for all $\boldsymbol{z} \in \mathcal{Z}_1$. Furthermore, we have the absolute continuity relation $P_{X,Y|Z=\boldsymbol{z}} \ll P_{X|Z=\boldsymbol{z}} \otimes P_{Y|Z=\boldsymbol{z}}$ with Radon-Nikodym derivative

$$\mathsf{S}_{\boldsymbol{z}} := \frac{\mathrm{d}P_{X,Y|Z=\boldsymbol{z}}}{\mathrm{d}(P_{X|Z=\boldsymbol{z}} \otimes P_{Y|Z=\boldsymbol{z}})}, \tag{51}$$

that satisfies $\mathbb{E}_{P_{X,Y|Z=\boldsymbol{z}}}[\mathsf{S}_{\boldsymbol{z}}(X,Y)] < +\infty$ for each $\boldsymbol{z} \in \mathcal{Z}_1$ and $\mathbb{E}_{P_{X,Y,Z}}[\mathsf{S}_Z(X,Y)] < +\infty$.

That $P_{X,Y,Z}$ marginalizes to $P_{X,Y}$ is more of an axiomatic property than an assumption, but we phrase it as so to emphasize that $P_{X,Y,Z}$ is meant to describe the evaluation distribution. The assumption that the conditionals $Q_{Z|X}$ and $P_{Z|X}$ match almost surely represents the viewpoint that, after fixing an image $\boldsymbol{x}$, the latent caption $Z|X = \boldsymbol{x}$ follows the same relationship to $\boldsymbol{x}$ as seen during pre-training. Importantly, this does not require or imply that $P_X = Q_X$ or that $P_Z = Q_Z$. The marginal distribution P_X is supplied entirely by the evaluation distribution $P_{X,Y}$, as for any measurable set $A \subseteq \mathcal{X}$, we have by definition that $P_X(A) = P_{X,Y}(A \times \mathcal{Y})$. On the other hand, the marginal P_Z can be defined using the Markov kernel $P_{Z|X=\boldsymbol{x}}$, in that for any measurable $B \subseteq \mathcal{Z}$, it holds that

$$P_Z(B) := \int_{\mathcal{X}_1} P_{Z|X=\boldsymbol{x}}(B)\mathrm{d}P_X(\boldsymbol{x}) = \int_{\mathcal{X}_1} Q_{Z|X=\boldsymbol{x}}(B)\mathrm{d}P_X(\boldsymbol{x}).$$

Finally, the absolute continuity condition, i.e., the existence of (51), rules out degeneracies such as Y being a deterministic function of X given $Z = \boldsymbol{z}$ (outside of a set of P_Z-measure zero). It is also worth pointing out that the first two conditions Asm. 13 do not contradict one another. For example, one can consider $P_{X,Y,Z}$ that satisfies the Markov chain $Y \to X \to Z$, where (X, Y) is drawn according to $P_{X,Y}$, and Z and Y are conditionally independent given X. Then, informally, we have that $P_{Z|X,Y} = P_{Z|X} = Q_{Z|X}$, so $P_{X,Y,Z}$ is uniquely determined. While this example implies the existence of a valid joint probability measure $P_{X,Y,Z}$, it is also, in a sense, showcasing the "least desirable" distribution for zero-shot prediction, as the dependence between X and Z does not provide any additional information about Y.

We recall some notation from Sec. 3. Let

$$g_{P_{Y,Z}}(\boldsymbol{z}) = \mathbb{E}_{P_{Y,Z}}[r(Y)|Z](\boldsymbol{z}).$$

Note that $g_{P_{Y,Z}}$ is simply a conditional expectation constructed via Definition 3, and does not require the existence of an r.c.d. $P_{Y|Z=\boldsymbol{z}}$. In the bound, we will encounter a prompt bias term that compares $g_{P_{Y,Z}}$ to g_ρ from (5). This reflects the notion that if $P_{X,Y,Z}$ agrees with two of the three fundamental distributions governing the problem, it will not be able to agree with the prompt distribution $\rho_{Y,Z}$ in general. Finally, the r.c.d. $P_{X,Y|Z=\boldsymbol{z}}$ allows us to measure conditional dependence using the *conditional mean squared contingency*, defined by the formula

$$I(X;Y|Z=\boldsymbol{z}) = \mathbb{E}_{P_{X|Z=\boldsymbol{z}} \otimes P_{Y|Z=\boldsymbol{z}}}\left[(1 - \mathsf{S}_{\boldsymbol{z}}(Y,X))^2\right].$$

As is shown in the proof, $I(X;Y|Z=\boldsymbol{z})$ and its expectation over P_Z are well-defined under Asm. 13. We are now ready to state the result.

Theorem 9. *Assume that r is bounded in absolute value by B_r. Under Asm. 13, it holds that Then, it holds that*

$$\|\eta_\rho - \eta_\star\|^2_{\mathbf{L}^2(P_X)} \le 2 \underbrace{\|g_\rho - g_{P_{Y,Z}}\|^2_{\mathbf{L}^2(P_Z)}}_{\text{prompt bias}} + 2B_r^2 \underbrace{\mathbb{E}_{P_Z}\left[I(X;Y|Z)\right]}_{\text{residual dependence}}. \tag{52}$$

Proof. We first establish a useful representation of the conditional mean of $r(Y)$ given $X = \boldsymbol{x}$, in terms of the (conditional) information density from Lem. 5. Fix $\boldsymbol{x} \in \mathcal{X}_1$ and $\boldsymbol{z} \in \mathcal{Z}_1$, the sets on which the regular conditional distributions $P_{Z|X=\boldsymbol{x}}$ and $P_{X,Y|Z=\boldsymbol{z}}$ are defined (see Asm. 13). Because of the existence the Radon-Nikodym derivative $\mathsf{S}_{\boldsymbol{z}}$ from (51), we may apply Lem. 5 with $U = Y$, $V = X$, and $h = r$ to write

$$\mathbb{E}_{P_{X,Y|Z=\boldsymbol{z}}}\left[r(Y)|X\right](\boldsymbol{x}) = \underbrace{\mathbb{E}_{P_{Y|Z=\boldsymbol{z}}}\left[r(Y)\mathsf{S}_{\boldsymbol{z}}(Y,\boldsymbol{x})\right]}_{=:f(\boldsymbol{x},\boldsymbol{z})} \text{ for all } (\boldsymbol{x},\boldsymbol{z}) \in \mathcal{X}_1 \times \mathcal{Z}_1.$$

The chosen notation $\mathbb{E}_{P_{X,Y|Z=\boldsymbol{z}}}\left[r(Y)|X\right](\boldsymbol{x})$ indicates that after fixing the probability measure $P_{X,Y|Z=\boldsymbol{z}}$, we take the conditional expectation of the function $r \in \mathbf{L}^2(P_{Y|Z=\boldsymbol{z}})$ via Definition 3, which does not necessarily posit the existence of the r.c.d. $P_{Y|X=\boldsymbol{x},Z=\boldsymbol{z}}$.[4] We have denoted the right-hand side by the function $f(\boldsymbol{x},\boldsymbol{z})$. Integrate both sides over $P_{Z|X=\boldsymbol{x}}$, then use the tower property of conditional expectation (Lem. 2) to achieve

$$\begin{aligned} \eta_\star(\boldsymbol{x}) = \mathbb{E}_{P_{X,Y}}\left[r(Y)|X\right](\boldsymbol{x}) &= \int_{\mathcal{Z}} \mathbb{E}_{P_{X,Y|Z=\boldsymbol{z}}}\left[r(Y)|X\right](\boldsymbol{x})\mathrm{d}P_{Z|X=\boldsymbol{x}}(\boldsymbol{z}) \\ &= \int_{\mathcal{Z}} f(\boldsymbol{x},\boldsymbol{z})\mathrm{d}P_{Z|X=\boldsymbol{x}}(\boldsymbol{z}) \\ &= \mathbb{E}_{P_{Z|X=\boldsymbol{x}}}\left[f(\boldsymbol{x},Z)\right]. \end{aligned} \tag{53}$$

Using the identity (53) and $Q_{Z|X=\boldsymbol{x}} = P_{Z|X=\boldsymbol{x}}$ on $\boldsymbol{x} \in \mathcal{X}_1$ (Asm. 13), we write

$$\begin{aligned} &\eta_\rho(\boldsymbol{x}) - \eta_\star(\boldsymbol{x}) \\ &= \mathbb{E}_{Q_{Z|X=\boldsymbol{x}}}\left[g_\rho(Z)\right] - \mathbb{E}_{P_{Z|X=\boldsymbol{x}}}\left[f(\boldsymbol{x},Z)\right] \\ &= \mathbb{E}_{P_{Z|X=\boldsymbol{x}}}\left[g_\rho(Z)\right] - \mathbb{E}_{P_{Z|X=\boldsymbol{x}}}\left[f(\boldsymbol{x},Z)\right] \\ &= \mathbb{E}_{P_{Z|X=\boldsymbol{x}}}\left[\left(g_\rho(Z) - g_{P_{Y,Z}}(Z)\right)\right] + \mathbb{E}_{P_{Z|X=\boldsymbol{x}}}\left[\left(g_{P_{Y,Z}}(Z) - f(\boldsymbol{x},Z)\right)\right]. \end{aligned}$$

Taking the integral over P_X, we have by the decomposition above that

$$\begin{aligned} \|\eta_\rho - \eta_\star\|^2_{\mathbf{L}^2(P_X)} &= \int_{\mathcal{X}} \left(\eta_\rho(\boldsymbol{x}) - \eta_\star(\boldsymbol{x})\right)^2 \mathrm{d}P_X(\boldsymbol{x}) \\ &\le 2 \int_{\mathcal{X}_1} \left(\mathbb{E}_{P_{Z|X=\boldsymbol{x}}}\left[g_\rho(Z) - g_{P_{Y,Z}}(Z)\right]\right)^2 \mathrm{d}P_X(\boldsymbol{x}) \tag{54} \\ &\quad + 2 \int_{\mathcal{X}_1} \left(\mathbb{E}_{P_{Z|X=\boldsymbol{x}}}\left[g_{P_{Y,Z}}(Z) - f(\boldsymbol{x},Z)\right]\right)^2 \mathrm{d}P_X(\boldsymbol{x}). \tag{55} \end{aligned}$$

To handle (54), we apply Jensen's inequality for each r.c.d. $P_{Z|X=\boldsymbol{x}}$ to achieve

$$\begin{aligned} \int_{\mathcal{X}_1} \left(\mathbb{E}_{P_{Z|X=\boldsymbol{x}}}\left[\left(g_\rho(Z) - g_{P_{Y,Z}}(Z)\right)\right]\right)^2 \mathrm{d}P_X(\boldsymbol{x}) &\le \int_{\mathcal{X}_1} \mathbb{E}_{P_{Z|X=\boldsymbol{x}}}\left[\left(g_\rho(Z) - g_{P_{Y,Z}}(Z)\right)^2\right] \mathrm{d}P_X(\boldsymbol{x}) \\ &= \mathbb{E}_{P_Z}\left[\left(g_\rho(Z) - g_{P_{Y,Z}}(Z)\right)^2\right] \\ &= \|g_\rho - g_{P_{Y,Z}}\|^2_{\mathbf{L}^2(P_Z)}. \end{aligned}$$

---

[4]This is why we do not write, for instance, $\mathbb{E}_{P_{Y|X=\boldsymbol{x},Z=\boldsymbol{z}}}\left[r(Y)\right]$. This consideration is purely technical, and the reader may make this substitution for conceptual understanding of the proof.

It remains to control (55). Applying Jensen's inequality for each r.c.d. $P_{Z|X=x}$ once again, we have that

$$\int_{\mathcal{X}_1} \left(\mathbb{E}_{P_{Z|X=x}}\left[\left(g_{P_{Y,Z}}(Z) - f(x,Z)\right)\right]\right)^2 \mathrm{d}P_X(x) \le \int_{\mathcal{X}_1} \left(\mathbb{E}_{P_{Z|X=x}}\left[\left(g_{P_{Y,Z}}(Z) - f(x,Z)\right)^2\right]\right) \mathrm{d}P_X(x)$$

$$= \mathbb{E}_{P_{X,Z}}\left[\left(g_{P_{Y,Z}}(Z) - f(X,Z)\right)^2\right]$$

$$= \int_{\mathcal{Z}_1} \mathbb{E}_{P_{X|Z=z}}\left[\left(g_{P_{Y,Z}}(z) - f(X,z)\right)^2\right] \mathrm{d}P_Z(z), \qquad (56)$$

where the last step follows due to the existence of the r.c.d. $P_{X|Z=z}$ for $z \in \mathcal{Z}_1$, as $P_{X|Z=z}(A) := P_{X,Y|Z=z}(A \times \mathcal{Y})$ for every measurable $A \subseteq \mathcal{X}$, and the latter exists by assumption. Using the definition of $g_{P_{Y,Z}}$, write

$$g_{P_{Y,Z}}(z) - f(x,z) = \mathbb{E}_{P_{Y|Z=z}}\left[r(Y)(1 - \mathsf{S}_z(Y,x))\right].$$

We may substitute this expression into the integrand of (56) and apply Jensen's inequality to $P_{Y|Z=z}$ to achieve

$$\mathbb{E}_{P_{X|Z=z}}\left[\left(g_{P_{Y,Z}}(z) - f(X,z)\right)^2\right] = \mathbb{E}_{P_{X|Z=z}}\left[\left(\mathbb{E}_{P_{Y|Z=z}}\left[r(Y)(1 - \mathsf{S}_z(Y,X))\right]\right)^2\right]$$

$$\le \mathbb{E}_{P_{X|Z=z}}\left[\mathbb{E}_{P_{Y|Z=z}}\left[\left(r(Y)(1 - \mathsf{S}_z(Y,X))\right)^2\right]\right]$$

$$\le \|r\|_\infty^2 \mathbb{E}_{P_{X|Z=z}}\left[\mathbb{E}_{P_{Y|Z=z}}\left[\left(1 - \mathsf{S}_z(Y,X)\right)^2\right]\right]$$

$$= \|r\|_\infty^2 \mathbb{E}_{P_{X|Z=z} \otimes P_{Y|Z=z}}\left[\left(1 - \mathsf{S}_z(Y,X)\right)^2\right],$$

where the final step follows by applying Fubini's theorem (Schilling, 2017, Corollary 14.9) to the product measure $P_{X|Z=z} \otimes P_{Y|Z=z}$ for fixed $z \in \mathcal{Z}_1$. By the definition of mean squared contingency (Definition 8), it holds that

$$\mathbb{E}_{P_{X|Z=z} \otimes P_{Y|Z=z}}\left[\left(1 - \mathsf{S}_z(Y,X)\right)^2\right] = I(X;Y|Z=z). \qquad (57)$$

After confirming that (57) is P_Z-integrable, substituting this expression back into (56) achieves the desired result. Expand the quadratic term and apply the Radon-Nikodym theorem (Schilling, 2017, Theorem 20.1) to achieve

$$I(X;Y|Z=z) = 1 - 2\mathbb{E}_{P_{X|Z=z} \otimes P_{Y|Z=z}}\left[\mathsf{S}_z(Y,X)\right] + \mathbb{E}_{P_{X|Z=z} \otimes P_{Y|Z=z}}\left[\mathsf{S}_z^2(Y,X)\right]$$

$$= 1 - 2\mathbb{E}_{P_{X,Y|Z=z}}[1] + \mathbb{E}_{P_{X,Y|Z=z}}\left[\mathsf{S}_z(Y,X)\right]$$

$$= \mathbb{E}_{P_{X,Y|Z=z}}\left[\mathsf{S}_z(Y,X)\right] - 1.$$

Thus, by integrating against P_Z, we see that

$$\mathbb{E}_{P_Z}\left[I(X;Y|Z)\right] = \mathbb{E}_{P_{X,Y,Z}}\left[\mathsf{S}_Z(Y,X)\right] - 1,$$

where the expectation term is finite by Asm. 13. The proof is complete. $\qquad\square$

D. Sample Complexity and Distribution Mismatch

This appendix provides the proofs of Thm. 2 and Thm. 3 by way of Thm. 10 and Thm. 11, respectively. To recall the bigger picture, we first applied the decomposition (12), which exposed the estimation error term

$$\|\hat{\eta}_\rho - \eta_\rho\|_{\mathbf{L}^2(P_X)}^2, \qquad (58)$$

where P_X is the $\mathcal{X}$-marginal of the evaluation distribution $P_{X,Y}$, η_ρ is defined by $\eta_\rho(x) := \mathbb{E}_{Q_{X,Z}}\left[g_\rho(Z)|X\right](x)$ (see (5)), and $\hat{\eta}_\rho$ is one of two estimation procedures that is based on either (7) or (8). By using standard change of measure arguments (collected in Appx. D.3), we pass the problem of controlling (58) in high probability to controlling $\|\hat{\eta}_\rho - \eta_\rho\|_{\mathbf{L}^2(Q_X)}^2$ (i.e. the mean squared error with respect to the pre-training marginal Q_X). Thus, the format of both Thm. 10 and Thm. 11 will be an upper bound on $\|\hat{\eta}_\rho - \eta_\rho\|_{\mathbf{L}^2(Q_X)}^2$ that holds with an arbitrary failure probability $\delta \in (0,1]$.

The identities (7) and (8) from Sec. 2 can be summarized with the equality

$$\mathbf{M}_{Z|X} g_\rho = \eta_\rho = \mathbb{E}_{\rho_{Y,Z}} \left[r(Y) \mathsf{R}(\cdot, Z) \right] + \mathrm{err}(Q_Z, \rho_Z), \tag{59}$$

where the $\mathrm{err}(Q_Z, \rho_Z)$ is elaborated on in Appx. D.2. In Appx. D.1, we consider the left-hand side of (59), and define $\hat{\eta}_\rho$ by constructing an estimate $\widehat{\mathbf{M}}_{Z|X}$ of $\mathbf{M}_{Z|X}$ using pre-training data and $\hat{g}_\rho$ of g_ρ using prompts. This will be referred to as the conditional mean approach. In Appx. D.2, we consider the right-hand side of (59) and define $\hat{\eta}_\rho$ by using an estimate $\widehat{\mathsf{R}}$ of R using pre-training data and $\hat{\rho}_{Y,Z}$ of $\rho_{Y,Z}$ using prompts. This will be referred to as the information density approach. For both approaches, we adopt a parallel structure and break the analysis into the following steps.

1. **Decomposing the Global Error:** We first provide a generic upper bound on the mean squared error

$$\|\hat{\eta}_\rho - \eta_\rho\|^2_{\mathbf{L}^2(Q_X)} \tag{60}$$

   in terms of the individual estimators defined by the pre-training and prompting data. While some additional structure may be employed in these bounds (e.g. the estimate lives in a reproducing kernel Hilbert space), the decomposition is generally agnostic to the choice of method and can accommodate multiple estimation/learning strategies.

2. **Interpreting the Source Condition:** The error term related to the pre-training data refers to the distance between the conditional mean operators $\widehat{\mathbf{M}}_{Z|X}$ and $\mathbf{M}_{Z|X}$ or the information densities $\widehat{\mathsf{R}}$ and R measured in an appropriate sense. This is initially controlled by substituting a particular estimation method among those reviewed in Appx. B. As mentioned in Sec. 3, the convergence rates of these methods rely on *source conditions* that describe the regularity of the target function. We derive expressions that relate the source conditions to measures of dependence between X and Z, so that, in turn, the rate can also be expressed in terms of these fundamental quantities.

3. **Controlling the Prompting Term:** The error term related to the prompting data will have a high probability bound, which is stated in the form of an assumption. This generality is maintained because the estimation based on the prompting data usually relies on simple primitives such as real-valued regression or finite-dimensional parameter estimation. Statistically, these problems are easier than the vector-valued regression or Radon-Nikodym derivative estimation problems that arise in the pre-training step. Thus, many possible methods can be used, and we provide examples in each case.

4. **Completing the Proof:** We combine the steps above to state the final bounds on (60). They are stated in Thm. 10 and Thm. 11, respectively.

The steps above comprise the subsections of Appx. D.1 and Appx. D.2 below. The bounds on mean square error on Q_X are tied to misclassification risk on $P_{X,Y}$ via Appx. D.3 and Appx. D.4 to produce end-to-end performance guarantees. We compare the sampling schemes used for prompting that are employed in the theoretical analysis to the sampling schemes used empirically in Appx. D.5.

D.1. Conditional Mean Approach

This approach is based on the LHS of (59) yielding the result of Thm. 2. The exposition relies heavily on the background introduced in Appx. B.4. In particular, we maintain the reproducing kernel Hilbert spaces $\mathcal{H}$ and $\mathcal{G}$ containing real-valued functions on $\mathcal{X}$ and $\mathcal{Z}$, respectively. We denote by $\mathbf{L}^2(Q_X; \mathcal{G})$ the Bochner space containing equivalence classes of functions mapping from $\mathcal{X}$ to $\mathcal{G}$. We also use the the bracket notation $[\cdot]_X$ to index a functions equivalence class in $\mathbf{L}^2(Q_X)$ (or $\mathbf{L}^2(Q_X; \mathcal{G})$ for $\mathcal{G}$-valued functions).

Setup. We first introduce an element $F_\star$ of $\mathbf{L}^2(Q_X; \mathcal{G})$ which can be used to represent the function $\boldsymbol{x} \mapsto [\mathbf{M}_{Z|X} g_\rho](\boldsymbol{x})$. We then describe how an estimator $\widehat{F}$ of $F_\star$ and an approximation $\hat{g}_\rho$ of g_ρ can be used to define an estimated predictor $\hat{\eta}_\rho$. Recall the boundedness assumptions $\sup_{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}} k(\boldsymbol{x}, \boldsymbol{x}') \leq k_{\max} < \infty$ and $\sup_{\boldsymbol{z}, \boldsymbol{z}' \in \mathcal{Z}} l(\boldsymbol{z}, \boldsymbol{z}') \leq l_{\max} < \infty$.

Lemma 8. *It holds that 1) $[\eta_\rho]_X \in \mathbf{L}^2(Q_X)$, and 2) there exists a function $F_\star : \mathcal{X} \to \mathcal{G}$ such that $[F_\star]_X \in \mathbf{L}^2(Q_X; \mathcal{G})$ and*

$$[\mathbb{E}_{Q_{X,Z}} [g(Z)|X] (\cdot)]_X = [\langle g, F_\star(\cdot) \rangle_{\mathcal{G}}]_X \text{ for all } g \in \mathcal{G}. \tag{61}$$

In particular, $[\eta_\rho]_X = [\langle g_\rho, F_\star(\cdot)\rangle_{\mathcal{G}}]_X$.

Proof. Using the notation from Appx. B.1, if we show that the random variable $\omega \mapsto g_\rho(Z(\omega))$ is contained in $L^2(\mathcal{F})$, then the first claim holds by the definition of conditional expectation in $\mathbf{L}^2(Q_X)$ (see Definition 3). Using the reproducing property of the RKHS $\mathcal{G}$, we have that

$$\mathbb{E}_{Q_Z}\left[g_\rho^2(Z)\right] = \mathbb{E}_{Q_Z}\left[\langle g_\rho, \psi(Z)\rangle_{\mathcal{G}}^2\right] \leq \|g_\rho\|_{\mathcal{G}}^2 \cdot \mathbb{E}_{Q_Z}\|\psi(Z)\|_{\mathcal{G}}^2 \leq l_{\max}\|g_\rho\|_{\mathcal{G}}^2, \tag{62}$$

granting the claim that $[\eta_\rho]_X \in \mathbf{L}^2(Q_X)$. Next, fix any $x \in \mathcal{X}$, and define the map

$$g \mapsto T_{\boldsymbol{x}}(g) = [\mathbf{M}_{Z|X}g](\boldsymbol{x}) = \mathbb{E}_{Q_{X,Z}}\left[g(Z)|X\right](\boldsymbol{x}).$$

By the same argument as (62), we have that $|T_{\boldsymbol{x}}(g)| \leq \sqrt{l_{\max}}\|g\|_{\mathcal{G}}$, indicating that $T_{\boldsymbol{x}}$ is a bounded linear functional. By the Riesz representation theorem, there exists an element of $\mathcal{G}$, denoted as $\mathbb{E}_{Q_{X,Z}}\left[\psi(Z)|X\right](\boldsymbol{x})$, that satisfies

$$\mathbb{E}_{Q_{X,Z}}\left[g(Z)|X\right](\boldsymbol{x}) = \left\langle g, \mathbb{E}_{Q_{X,Z}}\left[\psi(Z)|X\right](\boldsymbol{x})\right\rangle_{\mathcal{G}} \text{ for all } g \in \mathcal{G}.$$

Next, given the collection of Riesz representers $\left\{\mathbb{E}_{Q_{X,Z}}\left[\psi(Z)|X\right](\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}\right\}$, one may construct the mapping

$$F_\star : \mathcal{X} \to \mathcal{G}, \text{ defined by } \boldsymbol{x} \mapsto F_\star(\boldsymbol{x}) = \mathbb{E}_{Q_{X,Z}}\left[\psi(Z)|X\right](\boldsymbol{x}) \in \mathcal{G}.$$

It only remains to show that $[F_\star]_X \in \mathbf{L}^2(Q_X; \mathcal{G})$. By Jensen's inequality and the tower property (Lem. 2), we have that

$$\int_{\mathcal{X}}\|F_\star(\boldsymbol{x})\|_{\mathcal{G}}^2 \mathrm{d}Q_X(\boldsymbol{x}) \leq \mathbb{E}_{Q_Z}\|\psi(Z)\|_{\mathcal{G}}^2 \leq l_{\max} < \infty,$$

completing the proof. $\qquad\square$

Now that we have identified the vector-valued function of interest, $F_\star$, we can consider an estimation procedure that will return $\widehat{F} \equiv \widehat{F}_\lambda$, with $[\widehat{F}]_X \in \mathbf{L}^2(Q_X; \mathcal{G})$ and regularization parameter $\lambda > 0$. Then, we may define the estimator $\hat{\eta}_\rho$ of η_ρ via the inner product

$$\hat{\eta}_\rho(\boldsymbol{x}) = \langle \hat{g}_\rho, \widehat{F}(\boldsymbol{x})\rangle_{\mathcal{G}}, \tag{63}$$

where $\hat{g}_\rho$ satisfies some approximation bound with respect to g_ρ. Our decomposition will expose an error term for which we can apply Thm. 7, which describes the convergence rate of spectral regularization learning.

D.1.1. DECOMPOSING THE GLOBAL ERROR

Returning to the original quantity we wish to control from (58), we apply the following decomposition.

Lemma 9 (Error Decomposition). *For any choice of $\widehat{F} \in \mathbf{L}^2(Q_X; \mathcal{G})$ it holds that*

$$\|\hat{\eta}_\rho - \eta_\rho\|_{\mathbf{L}^2(Q_X)}^2 \leq 3\|g_\rho\|_{\mathcal{G}}^2 \cdot \|\widehat{F} - F_\star\|_{\mathbf{L}^2(Q_X; \mathcal{G})}^2 + 3\|F_\star\|_{\mathbf{L}^2(Q_X; \mathcal{G})}^2 \cdot \|\hat{g}_\rho - g_\rho\|_{\mathcal{G}}^2$$
$$+ 3\|\hat{g}_\rho - g_\rho\|_{\mathcal{G}}^2 \cdot \|\widehat{F} - F_\star\|_{\mathbf{L}^2(Q_X; \mathcal{G})}^2,$$

Proof. Using the reproducing property of the RKHS $\mathcal{G}$ and Young's inequality we have that

$$\|\hat{\eta}_\rho - \eta_\rho\|_{\mathbf{L}^2(Q_X)}^2$$
$$= \int_{\mathcal{X}}(\hat{\eta}_\rho(\boldsymbol{x}) - \eta_\rho(\boldsymbol{x}))^2 \mathrm{d}Q_X(\boldsymbol{x})$$
$$\leq 3\int_{\mathcal{X}}\langle g_\rho, \widehat{F}(\boldsymbol{x}) - F_\star(\boldsymbol{x})\rangle_{\mathcal{G}}^2 \mathrm{d}Q_X(\boldsymbol{x}) + 3\int_{\mathcal{X}}\langle \widehat{F}(\boldsymbol{x}), \hat{g}_\rho - g_\rho\rangle_{\mathcal{G}}^2 \mathrm{d}Q_X(\boldsymbol{x})$$
$$+ 3\int_{\mathcal{X}}\langle \widehat{F}(\boldsymbol{x}) - F_\star(\boldsymbol{x}), \hat{g}_\rho - g_\rho\rangle_{\mathcal{G}}^2 \mathrm{d}Q_X(\boldsymbol{x}).$$

Then, applying the Cauchy-Schwarz inequality in $\mathcal{G}$, we have that

$$
\begin{aligned}
&\|\hat{\eta}_\rho - \eta_\rho\|^2_{\mathbf{L}^2(Q_X)} \\
&\leq 3\|g_\rho\|^2_{\mathcal{G}} \cdot \int_{\mathcal{X}} \|\widehat{F}(\boldsymbol{x}) - F_\star(\boldsymbol{x})\|^2_{\mathcal{G}} \mathrm{d}Q_X(\boldsymbol{x}) + 3\left(\int_{\mathcal{X}} \|F_\star(\boldsymbol{x})\|^2_{\mathcal{G}} \mathrm{d}Q_X(\boldsymbol{x})\right) \cdot \|\hat{g}_\rho - g_\rho\|^2_{\mathcal{G}} \\
&\quad + 3\|\hat{g}_\rho - g_\rho\|^2_{\mathcal{G}} \int_{\mathcal{X}} \|\widehat{F}(\boldsymbol{x}) - F_\star(\boldsymbol{x})\|^2_{\mathcal{G}} \mathrm{d}Q_X(\boldsymbol{x}) \\
&= 3\|g_\rho\|^2_{\mathcal{G}} \cdot \|\widehat{F} - F_\star\|^2_{\mathbf{L}^2(Q_X;\mathcal{G})} + 3\|F_\star\|^2_{\mathbf{L}^2(Q_X;\mathcal{G})} \cdot \|\hat{g}_\rho - g_\rho\|^2_{\mathcal{G}} \\
&\quad + 3\|\hat{g}_\rho - g_\rho\|^2_{\mathcal{G}} \cdot \|\widehat{F} - F_\star\|^2_{\mathbf{L}^2(Q_X;\mathcal{G})},
\end{aligned}
\tag{64}
$$

the result as desired. $\qquad\square$

In the decomposition of Lem. 9, we observe the dominating terms $\|g_\rho\|^2_{\mathcal{G}} \cdot \|\widehat{F} - F_\star\|^2_{\mathbf{L}^2(Q_X;\mathcal{G})}$ and $\|F_\star\|^2_{\mathbf{L}^2(Q_X;\mathcal{G})} \cdot \|\hat{g}_\rho - g_\rho\|^2_{\mathcal{G}}$, along with the higher order term $\|\hat{g}_\rho - g_\rho\|^2_{\mathcal{G}} \cdot \|\widehat{F} - F_\star\|^2_{\mathbf{L}^2(Q_X;\mathcal{G})}$. We consider estimators $\widehat{F}$ and $\hat{g}_\rho$ based on kernel regularized learning techniques in order to bound the dominating terms, as a function of N and M. The bounds are optimized individually with respect to the regularization parameters of each learning objective.

D.1.2. INTERPRETING THE SOURCE CONDITION

To approach this, we associate our function of interest $F_\star \in \mathbf{L}^2(Q_X; \mathcal{G})$ to an object $\mathbf{C}_\star \in \mathrm{HS}(\mathbf{L}^2(Q_X), \mathcal{G})$ by way of an isometric isomorphism introduced in Thm. 6. This then allows us to derive a convenient formula for the quantity $\|F_\star\|_\beta$, which appears in Asm. 10, and relies on the interplay between $\mathcal{H}$ and $\mathbf{L}^2(Q_X)$ described in Appx. B.4.

Lemma 10. *Let $(g_j)_{j \in J}$ be any orthonormal basis (ONB) of $\mathcal{G}$ and recall the eigenfunctions $([e_{X,i}]_X)_{i \in I}$ from (36). Assuming that $\|F_\star\|_\beta$ is finite, it holds that*

$$
\|F_\star\|^2_\beta = \sum_{i \in I} \sum_{j \in J} \mu_{X,i}^{-\beta} \langle \mathbf{M}_{Z|X}[g_j]_Z, [e_{X,i}]_X \rangle^2_{\mathbf{L}^2(Q_X)}.
$$

Proof. By the definition of $\|\cdot\|_\beta$, we have that

$$
\|F_\star\|_\beta = \|\mathbf{C}_\star\|_\beta = \|\mathbf{C}_\star \mathbf{T}_X^{-\beta/2}\|_{\mathrm{HS}(\mathbf{L}^2(Q_X),\mathcal{G})}
\tag{65}
$$

Then, notice that by the eigendecomposition (36), we have that

$$
\mathbf{T}_X^{-\beta/2}[f]_X = 0 \text{ for all } [f]_X \in (\mathrm{cl}(\mathrm{range}(\mathbf{I}_X)))^\perp
$$

Thus, when computing the (65), we may restrict $\mathrm{HS}(\mathbf{L}^2(Q_X), \mathcal{G})$ to $\mathrm{HS}(\mathrm{cl}(\mathrm{range}(\mathbf{I}_X)), \mathcal{G})$. This allows us to employ the eigenvectors $([e_{X,i}]_X)_{i \in I}$ as a basis of $\mathrm{cl}(\mathrm{range}(\mathbf{I}_X))$ when computing the norm. We have that

$$
\begin{aligned}
&\|F_\star\|^2_\beta \\
&= \|\mathbf{C}_\star \mathbf{T}_X^{-\beta/2}\|^2_{\mathrm{HS}(\mathrm{cl}(\mathrm{range}(\mathbf{I}_X)),\mathcal{G})} \\
&= \sum_{i \in I} \sum_{j \in J} \langle g_j, \mathbf{C}_\star \mathbf{T}_X^{-\beta/2}[e_{X,i}]_X \rangle^2_{\mathcal{G}} && \text{(by definition)} \\
&= \sum_{i \in I} \sum_{j \in J} \mu_{X,i}^{-\beta} \langle g_j, \mathbf{C}_\star[e_{X,i}]_X \rangle^2_{\mathcal{G}} && \text{(by (36))} \\
&= \sum_{i \in I} \sum_{j \in J} \mu_{X,i}^{-\beta} \langle \mathbf{C}_\star, g_j \otimes [e_{X,i}]_X \rangle^2_{\mathrm{HS}(\mathbf{L}^2(Q_X),\mathcal{G})} \\
&= \sum_{i \in I} \sum_{j \in J} \sum_{k \in J} \sum_{l \in J} \mu_{X,i}^{-\beta} \langle g_k \otimes [f_k]_X, g_j \otimes [e_{X,i}]_X \rangle_{\mathrm{HS}(\mathbf{L}^2(Q_X),\mathcal{G})} \cdot \langle g_l \otimes [f_l]_X, g_j \otimes [e_{X,i}]_X \rangle_{\mathrm{HS}(\mathbf{L}^2(Q_X),\mathcal{G})}, && \text{(Lem. 7)}
\end{aligned}
$$

where $f_k(\boldsymbol{x}) = \langle F_\star(\boldsymbol{x}), g_k \rangle_{\mathcal{G}} = \mathbb{E}_{Q_{X,z}}[g_k(Z)|X](\boldsymbol{x})$. Phrased in terms of the conditional mean operator $\mathbf{M}_{Z|X}$:

$\mathbf{L}^2(Q_Z) \to \mathbf{L}^2(Q_X)$, we have that

$$[f_k]_X = \mathbf{M}_{Z|X}[g_k]_Z.$$

Plugging this into the display above, we have that

$$\|F_\star\|_\beta^2$$
$$= \sum_{i \in I} \sum_{j \in J} \sum_{k \in J} \sum_{l \in J} \mu_{X,i}^{-\beta} \big\langle g_k \otimes (\mathbf{M}_{Z|X}[g_k]_Z), g_j \otimes [e_{X,i}]_X \big\rangle_{\mathrm{HS}(\mathbf{L}^2(Q_X),\mathcal{G})} \big\langle g_l \otimes (\mathbf{M}_{Z|X}[g_l]_Z), g_j \otimes [e_{X,i}]_X \big\rangle_{\mathrm{HS}(\mathbf{L}^2(Q_X),\mathcal{G})}$$
$$= \sum_{i \in I} \sum_{j \in J} \sum_{k \in J} \sum_{l \in J} \mu_{X,i}^{-\beta} \langle g_k, g_j \rangle_\mathcal{G} \langle g_l, g_j \rangle_\mathcal{G} \cdot \big\langle \mathbf{M}_{Z|X}[g_k]_Z, [e_{X,i}]_X \big\rangle_{\mathbf{L}^2(Q_X)} \big\langle \mathbf{M}_{Z|X}[g_l]_Z, [e_{X,i}]_X \big\rangle_{\mathbf{L}^2(Q_X)}$$
$$= \sum_{i \in I} \sum_{j \in J} \mu_{X,i}^{-\beta} \big\langle \mathbf{M}_{Z|X}[g_j]_Z, [e_{X,i}]_X \big\rangle_{\mathbf{L}^2(Q_X)}^2,$$

where the last step follows from the fact that $g_1, g_2, \ldots$ is an ONB of $\mathcal{G}$. This completes the proof. $\qquad\square$

It remains to select a choice of the collection $(g_j)_{j \in J}$. Note that $([g_j]_Z)_{j \in J}$ does not form an orthonormal system in $\mathbf{L}^2(Q_Z)$, due to the distortion of the embedding. However, by explicitly writing the embedding $\mathbf{I}_Z$ (analogous to $\mathbf{I}_X$ introduced in (35)), we can derive one. Consider the singular value decomposition

$$\mathbf{I}_Z = \sum_{k \in K} \mu_{Z,k}^{1/2} \left([e_{Z,k}]_Z \otimes (\mu_{Z,k}^{1/2} e_{Z,k}) \right), \tag{66}$$

which is analogous to the one introduced for $\mathbf{I}_X$ in (38). The index set K is smaller in cardinality that J, as the collection $(e_{Z,k})_{k \in K}$ forms an ONB of $\mathrm{null}(\mathbf{I}_Z)^\perp \subseteq \mathcal{G}$, whereas $(g_j)_{j \in J}$ should be an ONB for all of $\mathcal{G}$. Thus, we can expand the embedding $[g_j]_Z \in \mathbf{L}^2(Q_Z)$ into

$$[g_j]_Z = \mathbf{I}_Z g_j = \sum_{k \in K} \mu_{Z,k}^{1/2} \big\langle g_j, \mu_{Z,k}^{1/2} e_{Z,k} \big\rangle_\mathcal{G} [e_{Z,k}]_Z.$$

This decomposition allows us to simplify the equality in Lem. 10 further.

Proposition 3. *In the setting of Lem. 10, it holds that*

$$\|F_\star\|_\beta^2 = \sum_{i \in I} \sum_{j \in J} \mu_{X,i}^{-\beta} \mu_{Z,j} \big\langle \mathbf{M}_{Z|X}[e_{Z,j}]_Z, [e_{X,i}]_X \big\rangle_{\mathbf{L}^2(Q_X)}^2 \tag{67}$$

$$= \|\mathbf{T}_X^{-\beta/2} \mathbf{M}_{Z|X} \mathbf{T}_Z^{1/2}\|_{\mathrm{HS}(\mathbf{L}^2(Q_Z), \mathbf{L}^2(Q_X))}^2. \tag{68}$$

In particular, $\|F_\star\|_0^2 = \|F_\star\|_{\mathbf{L}^2(Q_X;\mathcal{G})}^2 = \|\mathbf{M}_{Z|X} \mathbf{T}_Z^{1/2}\|_{\mathrm{HS}(\mathbf{L}^2(Q_Z), \mathbf{L}^2(Q_X))}^2.$

Proof. The sequence of functions $(\mu_{Z,k}^{1/2} e_{Z,k})_{k \in K}$ form an ONB of $\mathrm{null}(\mathbf{I}_Z)^\perp \subseteq \mathcal{G}$. Because J indexes a basis of $\mathcal{G}$, we have that $K \subseteq J$. Then, we may complete $(\mu_{Z,k}^{1/2} e_{Z,k})_{k \in K}$ to form the basis $(g_j)_{j \in J}$ of $\mathcal{G}$, where $g_j = \mu_{Z,j}^{1/2} e_{Z,j}$ for all $j \in K$ and g_j is defined arbitrarily for $j \notin K$. Plug $(g_j)_{j \in J}$ into the right-hand side of the formula given in Lem. 10 gives (67), the first part of the claim.

For the second equality, we note that $([e_{X,i}]_X)_{i \in I}$ and $([e_{Z,j}]_Z)_{j \in J}$ form orthonormal bases of $\mathrm{cl}(\mathrm{range}(\mathbf{I}_X))$ and $\mathrm{cl}(\mathrm{range}(\mathbf{I}_Z))$, respectively. We complete them (using the index sets $\bar{I}$ and $\bar{J}$) to form (possibly uncountable) orthonormal bases of $\mathbf{L}^2(Q_X)$ and $\mathbf{L}^2(Q_Z)$. Then, by the definition of the Hilbert-Schmidt norm, it holds that

$$\|\mathbf{T}_X^{-\beta/2} \mathbf{M}_{Z|X} \mathbf{T}_Z^{1/2}\|_{\mathrm{HS}(\mathbf{L}^2(Q_Z), \mathbf{L}^2(Q_X))}^2 = \sum_{i \in \bar{I}} \sum_{j \in \bar{J}} \big\langle \mathbf{T}_X^{-\beta/2} \mathbf{M}_{Z|X} \mathbf{T}_Z^{1/2} [e_{Z,j}]_Z, [e_{X,i}]_X \big\rangle_{\mathbf{L}^2(Q_X)}^2$$
$$= \sum_{i \in I} \sum_{j \in J} \mu_{X,i}^{-\beta} \mu_{Z,j} \big\langle \mathbf{M}_{Z|X}[e_{Z,j}]_Z, [e_{X,i}]_X \big\rangle_{\mathbf{L}^2(Q_X)}^2,$$

where we used in the second line that $[e_{X,i}]_X \in \text{null}(\mathbf{T}_X^{-\beta/2})$ for $i \in \bar{I} \backslash I$ and $[e_{Z,j}]_Z \in \text{null}(\mathbf{T}_Z^{1/2})$ for $j \in \bar{J} \backslash J$. This gives the (68) and completes the proof. □

It remains to interpret the equality in Prop. 3 to complete the analysis.

D.1.3. CONTROLLING THE PROMPTING TERM

From the decomposition given in Lem. 9, the estimate $\hat{g}_\rho$ will be designed as to control the RKHS-norm error $\|\hat{g}_\rho - g_\rho\|_{\mathcal{G}}^2$. We phrase the assumption generically, but in a way that is reflective of the convergence rates seen in real-valued nonparametric regression. Recall the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ introduced in Appx. B.1.

Assumption 14. For constants $\delta \in (0, 1]$, $M \geq 1$, and $\omega_\rho \in (1/2, 1]$, there is an event $\mathcal{E}(\delta, M, \omega_\rho)$ that is independent of the pre-training data $(X_1, Z_1), \ldots, (X_N, Z_N)$, such that on $\mathcal{E}(\delta, M, \omega_\rho)$,

$$\|\hat{g}_\rho - g_\rho\|_{\mathcal{G}}^2 \leq CB_r^2 \, \text{plog}(1/\delta) M^{-\frac{2\omega_\rho - 1}{2\omega_\rho + 1}}. \tag{69}$$

for a constant C independent of δ and M. On $(\Omega, \mathcal{F}, \mathbb{P})$, the event $\mathcal{E}(\delta, M, \omega_\rho)$ occurs with probability at least $1 - \delta/2$.

The notation ω_ρ is chosen for the constant that determines the convergence rate, because it can be interpreted itself as a source condition constant for a real-valued nonparametric regression framework. Indeed, consider the case in which $\hat{g}_\rho$ is computed using kernel ridge regression with parameter λ. Via the proof of their Theorem 2, Smale and Zhou (2007) show that with probability at least $1 - \delta/2$,

$$\|\hat{g}_\rho - g_\rho\|_{\mathcal{G}} \leq C(\rho_{Y,Z}) \log(4/\delta) \Big[\underbrace{B_r M^{-1/2} \lambda^{-1}}_{\text{estimation}} + \underbrace{\lambda^{\omega_\rho - 1/2}}_{\text{approximation}} \Big], \tag{70}$$

where $C(\rho_{Y,Z})$ is a constant that depends on the prompting measure $\rho_{Y,Z}$ and the choice of kernel. Optimizing the bound yields $\lambda \equiv \lambda_M \sim M^{-1/(2\omega_\rho + 1)}$, which ultimately leads to the convergence rate (notice the square) in (69). We comment that the choice to control the error in $\hat{g}_\rho$ in $\mathcal{G}$-norm comes from the vector-valued regression framework, in which the output space of the target function always lies in $\mathbf{L}^2(Q_X; \mathcal{G})$. In isolation, the mean squared error of $\hat{g}_\rho$ can be controlled both in $\mathbf{L}^2(\rho_Z)$-norm as well as interpolation norms in between $\mathbf{L}^2(\rho_Z)$ and $\mathcal{G}$ (see Fischer and Steinwart (2020), for instance). Indeed, when applying the decomposition (70) in $\mathbf{L}^2(\rho_Z)$-norm, Smale and Zhou (2007, Lemma 3) show that the approximation error decays as λ^{ω_ρ} (instead of $\lambda^{\omega_\rho - 1/2}$). In this case, the optimum is achieved at $\lambda_M \sim M^{-1/(2\omega_\rho + 2)}$, so that $\|\hat{g}_\rho - g_\rho\|_{\mathbf{L}^2(\rho_Z)}^2$ enjoys a convergence rate of $M^{-\omega_\rho/(\omega_\rho + 1)}$.

D.1.4. COMPLETING THE PROOF

We may now prove Thm. 2. Next, we place the requisite conditions on β, given eigendecay assumptions on $\mathbf{T}_X$ and $\mathbf{T}_Z$, and singular decay assumptions on $\mathbf{M}_{Z|X}$ (see Appx. B.2 for a review of these operator decompositions). Under these assumptions, we will have that all operators will have a countably infinite number of non-zero eigenvalues/singular values.

Assumption 15 (Eigendecay and Singular Decay). Let the eigenvalues of $\mathbf{T}_X$, eigenvalues of $\mathbf{T}_Z$, and singular values of $\mathbf{M}_{Z|X}$ be given by $\{\mu_{X,i}\}_{i=1}^\infty$, $\{\mu_{Z,i}\}_{i=1}^\infty$, and $\{\sigma_i\}_{i=1}^\infty$, respectively. There exist positive constants c, C, γ_X, γ_Z, and $\gamma_{X,Z}$ such that for all $i = 1, 2, \ldots$, we have the inclusions

$$\mu_{X,i} \in \left[ci^{-\gamma_X}, Ci^{-\gamma_X} \right], \mu_{Z,i} \in \left[ci^{-\gamma_Z}, Ci^{-\gamma_Z} \right], \text{ and } \sigma_i \in \left[ci^{-\gamma_{X,Z}}, Ci^{-\gamma_{X,Z}} \right].$$

Assumption 16 (Basis Alignment). There exists a finite index $m \in \mathbb{N}$ and a permutation $\pi : [m] \to [m]$ such that the operator $\mathbf{M}_{Z|X}$ admits the singular value decomposition

$$\mathbf{M}_{Z|X} = \sum_{i=1}^m \sigma_{\pi(i)} [e_{Z,i}]_Z \otimes [e_{X,i}]_Z + \sum_{j=m+1}^\infty \sigma_i [e_{Z,i}]_Z \otimes [e_{X,i}]_Z.$$

Asm. 16 allows us to reason about the finiteness of the Hilbert-Schmidt norm $\|\mathbf{T}_X^{-\beta/2} \mathbf{M}_{Z|X} \mathbf{T}_Z^{1/2}\|_{\text{HS}(\mathbf{L}^2(Q_Z), \mathbf{L}^2(Q_X))}^2$ based on the eigendecays of the various operators introduced in Asm. 15. These will imply a maximal value of the source

condition constant β.

Lemma 11. *Under Asm. 15 and Asm. 16, it holds that $\|F_\star\|_\beta < +\infty$ if and only if*

$$\beta < \frac{2\gamma_{X,Z} + \gamma_Z - 1}{\gamma_X}. \tag{71}$$

Proof. For ease of presentation, we extend the permutation π from Asm. 16 so that $\pi(i) = i$ for all $i \geq m + 1$. Applying the result from Prop. 3, and using the eigenbases of $\mathbf{T}_X$ and $\mathbf{T}_Z$, we see that

$$
\begin{aligned}
\|F_\star\|_\beta^2 &= \|\mathbf{T}_X^{-\beta/2} \mathbf{M}_{Z|X} \mathbf{T}_Z^{1/2}\|_{\mathrm{HS}(\mathbf{L}^2(Q_Z), \mathbf{L}^2(Q_X))}^2 \\
&\geq \frac{c}{C^\beta} \sum_{i=1}^\infty \sum_{j=1}^\infty j^{-\gamma_Z} i^{\beta\gamma_X} \big\langle \mathbf{M}_{Z|X}[e_{Z,j}]_Z, [e_{X,i}]_X \big\rangle_{\mathbf{L}^2(Q_X)}^2 &\text{(Asm. 15)}\\
&= \frac{c}{C^\beta} \sum_{i=1}^\infty \sum_{j=1}^\infty \sum_{k=1}^\infty \sum_{l=1}^\infty j^{-\gamma_Z} i^{\beta\gamma_X} \sigma_{\pi(k)}^2 \langle [e_{Z,k}]_Z, [e_{Z,j}]_Z \rangle_{\mathbf{L}^2(Q_Z)} \langle [e_{X,k}]_X, [e_{X,i}]_X \rangle_{\mathbf{L}^2(Q_X)} \\
&\quad \times \langle [e_{Z,l}]_Z, [e_{Z,j}]_Z \rangle_{\mathbf{L}^2(Q_Z)} \langle [e_{X,l}]_X, [e_{X,i}]_X \rangle_{\mathbf{L}^2(Q_X)} \\
&= \frac{c}{C^\beta} \left[\sum_{i=1}^m i^{\beta\gamma_X - \gamma_Z} \sigma_{\pi(i)}^2 + \sum_{i=m+1}^\infty i^{\beta\gamma_X - \gamma_Z} \sigma_i^2 \right] &\text{(Asm. 16)}\\
&\geq \frac{c}{C^\beta} \left[\sum_{i=1}^m i^{\beta\gamma_X - \gamma_Z} \sigma_{\pi(i)}^2 + c \sum_{i=m+1}^\infty i^{\beta\gamma_X - \gamma_Z - 2\gamma_{X,Z}} \right], &\text{(Asm. 15)}
\end{aligned}
$$

where the rightmost term is finite only if (71) holds. Arguing similarly for the upper bound, we have that

$$\|F_\star\|_\beta^2 \leq \frac{C}{c^\beta} \left[\sum_{i=1}^m i^{\beta\gamma_X - \gamma_Z} \sigma_{\pi(i)}^2 + C \sum_{i=m+1}^\infty i^{\beta\gamma_X - \gamma_Z - 2\gamma_{X,Z}} \right],$$

where we may claim that $\|F_\star\|_\beta^2 < +\infty$ if (71) holds. $\qquad\square$

We can now wrap together the results of this section. Recalling the estimator $\widehat{F} \equiv \widehat{F}_\lambda$ based on vector-valued spectral regularization learning, described in Appx. B.4. The well-specified case refers to the condition that $\beta \geq 1$, indicating that the RKHS in which $\widehat{F}$ is learned does indeed contain $F_\star$. When $\beta < 1$, we require more sophisticated tools, namely, vector-valued interpolation spaces. In both cases, after establishing the results above, we capture the sample complexity via Thm. 7 from Appx. B.4.

Well-Specified Case. Under Asm. 15 and Asm. 16, this implies via Lem. 11 that

$$1 \leq \beta = \left(\frac{2\gamma_{X,Z} + \gamma_Z - 1}{\gamma_X} \right)^t < \frac{2\gamma_{X,Z} + \gamma_Z - 1}{\gamma_X}, \text{ for } t \in [0,1). \tag{72}$$

Thus, we may use the parameter $t \in [0,1)$ to measure the degree to which the upper bound is saturated. This yields the following result, which reflects Thm. 2 from the main text. To state the result, define the quantity

$$q(t) = (2\gamma_{X,Z} + \gamma_Z - 1)^t \gamma_X^{1-t} \tag{73}$$

and observe the following, which is an immediate consequence of Lem. 9, Thm. 7, and the formula (72). Note that the constant p in Thm. 7 refers to $1/\gamma_X$ in the notation of this section.

Theorem 10. *Consider failure probability $\delta \in (0,1]$. Let Asm. 14, Asm. 15, Asm. 16, and the conditions of Thm. 7 hold with $\|\mathbf{T}_X^{-1/2} \mathbf{M}_{Z|X} \mathbf{T}_Z^{1/2}\|_{\mathrm{HS}(\mathbf{L}^2(Q_Z), \mathbf{L}^2(Q_X))}^2 < +\infty$. Then, for $\hat{\eta}_\rho$ defined via (63), there exist a constants $t \in [0,1)$ and*

$C \geq 0$ *such that with probability at least* $1 - \delta$,

$$\|\hat{\eta}_\rho - \eta_\rho\|_{\mathbf{L}^2(Q_X)}^2 \lesssim \mathrm{plog}(1/\delta) \left[N^{-\frac{q(t)}{q(t)+1}} + B_r^2 \|\mathbf{M}_{Z|X} \mathbf{T}_Z^{1/2}\|_{\mathrm{HS}}^2 M^{-\frac{2\omega_\rho - 1}{2\omega_\rho + 1}} \right]$$

for all $N \geq C \,\mathrm{plog}(N/\delta)$. *where* $\|\cdot\|_{\mathrm{HS}} = \|\cdot\|_{\mathrm{HS}(\mathbf{L}^2(Q_Z), \mathbf{L}^2(Q_X))}$.

The term $\|\mathbf{M}_{Z|X} \mathbf{T}_Z^{1/2}\|_{\mathrm{HS}}^2$ is equal (via Prop. 3) to the $\|F_\star\|_{\mathbf{L}^2(Q_X; \mathcal{G})}^2$ term from Lem. 9, and is rendered (along with B_r^2 as the constant $C(Q_{X,Z})$ in Thm. 2.

Mis-Specified Case. The first inequality of (72) holds only when $F_\star$ is well-specified, or contained in the vector-valued RKHS used in the estimation procedure that defines (43). We may employ the interpolation space machinery from Appx. B.4 to achieve a convergence guarantee in this setting. Recall the constant $\alpha \in [1/\gamma_X, 1]$ shown in Asm. 10, which is associated to the continuous embedding $\mathbf{I}_X^{\alpha; \infty} : [\mathcal{H}]^\alpha \hookrightarrow \mathbf{L}^\infty(Q_X)$. This constant describes the RKHS itself, and not the specific target function $F_\star$. The rate of Thm. 10 may still be achieved for function classes that are "not too mis-specified" in the sense of Case 1 from Thm. 7. The inequality (71) provides a sufficient condition for Case 2, that is, when $\beta + 1/\gamma_X \leq \alpha$. Indeed,

$$\frac{2\gamma_{X,Z} + \gamma_Z}{\gamma_X} \leq \alpha \implies \beta + 1/\gamma_X \overset{(71)}{<} \frac{2\gamma_{X,Z} + \gamma_Z}{\gamma_X} \leq \alpha. \tag{74}$$

The left-hand side may also be phrased differently as $2\gamma_{X,Z} + \gamma_Z \leq \alpha \gamma_X$. Thus, we may interpret $\alpha \gamma_X \in [1, \gamma_X]$ as a parameter that controls the mis-specification threshold. Concretely, it becomes easier for $F_\star$ to be mis-specified when: $\gamma_{X,Z}$ is low ((X, Z) are highly dependent), γ_Z is low (the effective dimension of Z is large), or γ_X is high (the effective dimension of the input X is small). Under the sufficient condition (74), along with Asm. 15 and Asm. 16, the best upper bound on the convergence rate in the current mis-specification model (see Thm. 7, Case 2), is then

$$\|\hat{\eta}_\rho - \eta_\rho\|_{\mathbf{L}^2(Q_X)}^2 \lesssim \mathrm{plog}(1/\delta) \left[N^{-\frac{2\gamma_{X,Z} + \gamma_Z - 1}{\alpha \gamma_X}} + B_r^2 \|\mathbf{M}_{Z|X} \mathbf{T}_Z^{1/2}\|_{\mathrm{HS}}^2 M^{-\frac{2\omega_\rho - 1}{2\omega_\rho + 1}} \right]$$

for N sufficiently large.

D.2. Information Density Approach

This approach is based on the RHS of (59) and yields the result of Thm. 3. Here, we assume that during the pre-training phase, the user produces an estimated function $\widehat{R}$, which is an element of a reproducing kernel Hilbert space (RKHS). Unlike in Appx. D.1, where we approximated g_ρ using a function $\hat{g}_\rho$ (which aligns with the conditional mean viewpoint), the information density viewpoint in this section warrants estimating the mean of a function under $\rho_{Y,Z}$ directly, using samples $(Y_1, Z_1), \ldots, (Y_M, Z_M) \overset{\mathrm{i.i.d}}{\sim} \rho_Z$. It is also important to point out a slight difference in the sampling model for the pre-training data. In order to define the estimate (45) for our method of choice (and similar Radon-Nikodym derivative estimation techniques), it is typically assumed that we observe data from both distributions in the ratio. In the case of $Q_{X,Z}$ and $Q_X \otimes Q_Z$, this corresponds to observing N_{p} paired examples and N_{u} unpaired examples such that $N = N_{\mathrm{p}} + N_{\mathrm{u}}$. For simplicity, we assume that $N_{\mathrm{p}} = N_{\mathrm{u}} = N/2$, but remark that the regime in which $N_{\mathrm{u}} \gg N_{\mathrm{p}}$ is an interesting and practically relevant model for future investigations.

Setup. Let $\mathcal{S}$ denote a separable reproducing kernel Hilbert space (RKHS) of real-valued functions on $\mathcal{X} \times \mathcal{Z}$, with canonical feature map $\varphi : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ and reproducing kernel $\kappa : (\mathcal{X} \times \mathcal{Z}) \times (\mathcal{X} \times \mathcal{Z}) \to \mathbb{R}$. We will express the error in terms of the RKHS norm difference $\|\widehat{R} - R\|_{\mathcal{S}}^2$, among other terms that capture a notion of "distribution mismatch" between the prompting marginal ρ_Z and the pre-training marginal Q_Z. This may also be interpreted as another instance of prompt bias. This error occurs because at prompting time, the user does not necessarily have any data drawn from Q_Z. As before, we maintain $\sup \{\kappa(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{x}', \boldsymbol{z}') : (\boldsymbol{x}, \boldsymbol{z}), (\boldsymbol{x}', \boldsymbol{z}') \in \mathcal{X} \times \mathcal{Z}\} \leq \kappa_{\max}$.

Recall that the true R is a kernel for the conditional mean operator when integrated under Q_Z (see Lem. 5), but can also be

related via Lem. 6 to the marginal distribution ρ_Z:

$$\eta_\rho(\boldsymbol{x}) = \mathbb{E}_{\rho_{Y,Z}}\left[r(Y)\mathsf{R}(\boldsymbol{x}, Z)\right] + \int_{\mathcal{Z}} g_\rho(\boldsymbol{z})\mathsf{R}(\boldsymbol{x}, \boldsymbol{z})\left(\mathrm{d}Q_Z(\boldsymbol{z}) - \mathrm{d}\rho_Z(\boldsymbol{z})\right).$$

This motivates the approximation $\hat{\rho}_{Y,Z}$ expressed directly in terms of the prompt distribution, and the estimator

$$\hat{\eta}_\rho(\boldsymbol{x}) = \mathbb{E}_{\hat{\rho}_{Y,Z}}[r(Y)\widehat{\mathsf{R}}(\boldsymbol{x}, Z)]. \tag{75}$$

Below, we consider the empirical measure

$$\hat{\rho}_{Y,Z} = \frac{1}{M}\sum_{j=1}^{M}\delta_{(Y_j, Z_j)} \tag{76}$$

so that for fixed $\boldsymbol{x} \in \mathcal{X}$, (75) reduces to a sample mean.

D.2.1. DECOMPOSING THE GLOBAL ERROR

The estimation error decomposition below will take the two differences into account: between the marginal distributions Q_Z and ρ_Z and between the joint distribution $\hat{\rho}_{Y,Z}$ and $\rho_{Y,Z}$. For the latter, we will define random variables that take values in a Hilbert space (specifically, $\mathbf{L}^2(Q_X)$). This will allow for controlling deviations between $\hat{\rho}_{Y,Z}$ and $\rho_{Y,Z}$ directly for the test functions being integrated. Define the independent and identically random variables $W_1, \ldots, W_M$ by

$$W_j := r(Y_j)\mathsf{R}(\cdot, Z_j),$$

and the element of $\mathbf{L}^2(Q_X)$ (interpreted as the expectation) $\mathbb{E}_{\rho_{Y,Z}}[W_1] : \boldsymbol{x} \mapsto \mathbb{E}_{\rho_{Y,Z}}[r(Y_1)\mathsf{R}(\boldsymbol{x}, Z_1)]$.

Lemma 12 (Error Decomposition). *Assume the following conditions.*

- *$\rho_Z \ll Q_Z$ with Q_Z-square integrable Radon-Nikodym derivative (i.e. $\chi^2(\rho_Z\|Q_Z) < +\infty$).*

- *R is contained in $\mathbf{L}^2(Q_X \otimes \rho_Z)$ and $\mathbf{L}^2(Q_X \otimes Q_Z)$.*

Then, it holds that

$$\|\hat{\eta}_\rho - \eta_\rho\|_{\mathbf{L}^2(Q_X)}^2 \le 3B_r^2\left(\kappa_{\max}^2\|\widehat{\mathsf{R}} - \mathsf{R}\|_{\mathcal{S}}^2 + \|\mathsf{R}\|_{\mathbf{L}^2(Q_X \otimes Q_Z)}^2\chi^2(\rho_Z\|Q_Z)\right) \tag{77}$$
$$+ 3\|\tfrac{1}{M}\sum_{j=1}^{M}W_j - \mathbb{E}_{\rho_{Y,Z}}[W_1]\|_{\mathbf{L}^2(Q_X)}^2.$$

Proof. Using Lem. 6, we have that for Q_X-almost all $\boldsymbol{x} \in \mathcal{X}$,

$$\hat{\eta}_\rho(\boldsymbol{x}) - \eta_\rho(\boldsymbol{x}) = \mathbb{E}_{\hat{\rho}_{Y,Z}}[r(Y)\widehat{\mathsf{R}}(\boldsymbol{x}, Z)] - \mathbb{E}_{\rho_{Y,Z}}[r(Y)\mathsf{R}(\boldsymbol{x}, Z)]$$
$$+ \int_{\mathcal{Z}} g_\rho(\boldsymbol{z})\mathsf{R}(\boldsymbol{x}, \boldsymbol{z})\left(\mathrm{d}Q_Z(\boldsymbol{z}) - \mathrm{d}\rho_Z(\boldsymbol{z})\right)$$
$$= \mathbb{E}_{\hat{\rho}_{Y,Z}}[r(Y)\langle\varphi(\boldsymbol{x}, Z), \widehat{\mathsf{R}} - \mathsf{R}\rangle]$$
$$+ \int_{\mathcal{Y}\times\mathcal{Z}} r(\boldsymbol{y})\mathsf{R}(\boldsymbol{x}, \boldsymbol{z})\left(\mathrm{d}\hat{\rho}_{Y,Z}(\boldsymbol{y}, \boldsymbol{z}) - \mathrm{d}\rho_{Y,Z}(\boldsymbol{y}, \boldsymbol{z})\right)$$
$$+ \int_{\mathcal{Z}} g_\rho(\boldsymbol{z})\mathsf{R}(\boldsymbol{x}, \boldsymbol{z})\left(\mathrm{d}Q_Z(\boldsymbol{z}) - \mathrm{d}\rho_Z(\boldsymbol{z})\right).$$

42

Then, we have that

$$\|\hat{\eta}_\rho - \eta_\rho\|^2_{\mathbf{L}^2(Q_X)} \leq 3\mathbb{E}_{Q_X}\left[\left(\mathbb{E}_{\hat{\rho}_{Y,Z}}[r(Y)\langle\varphi(X,Z),\widehat{\mathsf{R}} - \mathsf{R}\rangle]^2\right)\right] \tag{78}$$

$$+ 3\int_{\mathcal{X}}\left(\int_{\mathcal{Y}\times\mathcal{Z}} r(\boldsymbol{y})\mathsf{R}(\boldsymbol{x},\boldsymbol{z})\left(\mathrm{d}\hat{\rho}_{Y,Z}(\boldsymbol{y},\boldsymbol{z}) - \mathrm{d}\rho_{Y,Z}(\boldsymbol{y},\boldsymbol{z})\right)\right)^2 \mathrm{d}Q_X(\boldsymbol{x}) \tag{79}$$

$$+ 3\int_{\mathcal{X}}\left(\int_{\mathcal{Z}} g_\rho(\boldsymbol{z})\mathsf{R}(\boldsymbol{x},\boldsymbol{z})\left(\mathrm{d}Q_Z(\boldsymbol{z}) - \mathrm{d}\rho_Z(\boldsymbol{z})\right)\right)^2 \mathrm{d}Q_X(\boldsymbol{x}). \tag{80}$$

To control (78), apply boundedness to achieve

$$\mathbb{E}_{Q_X}\left[\left(\mathbb{E}_{\hat{\rho}_{Y,Z}}[r(Y)\langle\varphi(X,Z),\widehat{\mathsf{R}} - \mathsf{R}\rangle]^2\right)\right] \leq B_r^2\kappa_{\max}^2\|\widehat{\mathsf{R}} - \mathsf{R}\|^2_{\mathcal{S}}.$$

For (79), the term is equal to $\|\frac{1}{M}\sum_{j=1}^M W_j - \mathbb{E}_{\rho_{Y,Z}}[W_1]\|^2_{\mathbf{L}^2(Q_X)}$ by definition of $W_1, \ldots, W_M$. For (80), we use that $\rho_Z \ll Q_Z$ and $\|g_\rho\|_\infty \leq B_r$ and apply the Cauchy-Schwarz inequality on $\mathbf{L}^2(Q_Z)$ so that

$$\left(\int_{\mathcal{Z}} g_\rho(\boldsymbol{z})\mathsf{R}(\boldsymbol{x},\boldsymbol{z})\left(\mathrm{d}Q_Z(\boldsymbol{z}) - \mathrm{d}\rho_Z(\boldsymbol{z})\right)\right)^2$$

$$= \left(\int_{\mathcal{Z}} g_\rho(\boldsymbol{z})\mathsf{R}(\boldsymbol{x},\boldsymbol{z})\left(1 - \frac{\mathrm{d}\rho_Z}{\mathrm{d}Q_Z}(\boldsymbol{z})\right)\mathrm{d}Q_Z(\boldsymbol{z})\right)^2$$

$$\leq \|r\|^2\|\mathsf{R}(\boldsymbol{x},\cdot)\|^2_{\mathbf{L}^2(Q_Z)}\underbrace{\int_Z\left(1 - \frac{\mathrm{d}\rho_Z}{\mathrm{d}Q_Z}(\boldsymbol{z})\right)^2 \mathrm{d}Q_Z(\boldsymbol{z})}_{\chi^2(\rho_Z\|Q_Z)}.$$

Taking the expectation over Q_X gives $\mathbb{E}_{Q_X}\|\mathsf{R}(X,\cdot)\|^2_{\mathbf{L}^2(Q_Z)} = \|\mathsf{R}\|^2_{\mathbf{L}^2(Q_X\otimes Q_Z)}$ and completes the proof. $\qquad\square$

Given the decomposition shown in Lem. 12, it remains to bound both the error term $\|\widehat{\mathsf{R}} - \mathsf{R}\|^2_{\mathcal{S}}$ regarding the estimated Radon-Nikodym derivative $\widehat{\mathsf{R}}$, and the approximation term $\|\frac{1}{M}\sum_{j=1}^M W_j - \mathbb{E}_{\rho_{Y,Z}}[W_1]\|^2_{\mathbf{L}^2(Q_X)}$. We will employ Cor. 2 to this end. Unlike the arguments of Appx. D.1, there is only a single kernel regularized learning algorithm at play, that is, for the estimation of $\widehat{\mathsf{R}}$. We proceed to interpret the source condition Asm. 12.

D.2.2. INTERPRETING THE SOURCE CONDITION

To proceed, we introduce some notation related to $\mathbf{L}^2(Q_X \otimes Q_Z)$ and the RKHS $\mathcal{S}$. These objects are also introduced in Appx. B.4, so we review their properties briefly. Let $[h]_\sim$ index the equivalence class in $\mathbf{L}^2(Q_X \otimes Q_Z)$ for a square-integrable function $h : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$. This indexing can also be identified with an *embedding operator* $\mathbf{I}_{X,Z} : \mathcal{S} \to \mathbf{L}^2(Q_X \otimes Q_Z)$, which is Hilbert-Schmidt under the boundedness of the kernel κ by $\kappa_{\max}$. Letting $\mathbf{S}_{X,Z} = \mathbf{I}^*_{X,Z} : \mathbf{L}^2(Q_X \otimes Q_Z) \to \mathcal{S}$ be its adjoint, we have that $\mathbf{I}_{X,Z}\mathbf{S}_{X,Z} : \mathbf{L}^2(Q_X \otimes Q_Z) \to \mathbf{L}^2(Q_X \otimes Q_Z)$ and $\mathbf{S}_{X,Z}\mathbf{I}_{X,Z} : \mathcal{S} \to \mathcal{S}$ are compact, trace class operators. These form the analogs of $(\mathbf{T}_X, \mathbf{T}_Z)$ and $(\mathbf{C}_X, \mathbf{C}_Z)$, respectively, from Appx. B.4. Via Thm. 4, we write the eigendecomposition

$$\mathbf{I}_{X,Z}\mathbf{S}_{X,Z} = \sum_{i\in I}\mu_i\langle\cdot,[e_i]_\sim\rangle_{\mathbf{L}^2(Q_X\otimes Q_Z)}[e_i]_\sim, \tag{81}$$

where we may take each representative e_i as an element of $\mathcal{S}$ (Steinwart and Scovel, 2012, Lemma 2.12). Then, we also have that

$$\mathbf{S}_{X,Z}\mathbf{I}_{X,Z} = \sum_{i\in I}\mu_i\langle\cdot,\mu_i^{1/2}e_i\rangle_{\mathcal{S}}\mu_i^{1/2}e_i, \tag{82}$$

These constructions (along with Prop. 2) give us the following relationship between the Hilbert-Schmidt norm of the conditional mean operator $\mathbf{M}_{Z|X}$ and the Radon-Nikodym derivative under the condition Asm. 12. In fact, finiteness follows from the source condition itself and boundedness of the kernel.

Lemma 13. *Under Asm. 12, it holds that*

$$\|\mathbf{M}_{Z|X}\|^2_{\mathrm{HS}(\mathbf{L}^2(Q_Z),\mathbf{L}^2(Q_X))} = \|\mathbf{I}_{X,Z}\mathsf{R}\|^2_{\mathbf{L}^2(Q_X \otimes Q_Z)} = \sum_{i \in I} \mu_i^{2\beta+1} \langle \mathsf{S}_{Q_{X,Z}}, \mu_i^{1/2} e_i \rangle^2_{\mathcal{S}}.$$

Proof. Without loss of generality, assume that $\mathsf{R} \in \mathrm{null}(\mathbf{I}_{X,Z})^\top$ (as the component in $\mathrm{null}(\mathbf{I}_{X,Z})$ will be excluded from the norm calculation anyway). We expand the expression for R appearing in Asm. 12 on an ONB of $\mathrm{null}(\mathbf{I}_{X,Z})^\top$. To do so, combine (81) and (82) to introduce the singular value decomposition

$$\mathbf{I}_{X,Z} = \sum_{i \in I} \mu_i^{1/2} \langle \cdot, \mu_i^{1/2} e_i \rangle_{\mathcal{S}} [e_i]_\sim.$$

Then, it holds under Asm. 12 that

$$\mathsf{R} = \sum_{i \in I} \mu_i^{\beta} \langle \mathsf{S}_{Q_{X,Z}}, \mu_i^{1/2} e_i \rangle_{\mathcal{S}} \mu_i^{1/2} e_i \text{ and } \mathbf{I}_{X,Z}\mathsf{R} = \sum_{i \in I} \mu_i^{\beta+1/2} \langle \mathsf{S}_{Q_{X,Z}}, \mu_i^{1/2} e_i \rangle_{\mathcal{S}} [e_i]_\sim.$$

Using that $([e_i]_\sim)_{i \in I}$ is an orthonormal system, we may use the second expression to perform the computation. $\qquad\square$

To make use of Lem. 13, we now interpret β in terms of eigendecay exponents of the operators in question.

Assumption 17 (Eigendecay and Singular Decay). Let the eigenvalues of $\mathbf{I}_{X,Z}\mathbf{S}_{X,Z}$ and singular values of $\mathbf{M}_{Z|X}$ be given by $\{\mu_i\}_{i=1}^\infty$ and $\{\sigma_i\}_{i=1}^\infty$, respectively. There exist positive constants $c, C, \alpha > 1$, and $\gamma_{X,Z} > 1/2$ such that for all $i = 1, 2, \ldots$, we have the inclusions

$$\mu_i \leq \left[ci^{-\alpha}, Ci^{-\alpha} \right]. \text{ and } \sigma_i \in \left[ci^{-\gamma_{X,Z}}, Ci^{-\gamma_{X,Z}} \right].$$

The following relationship holds over an interval in β. We explicitly account for the dependence of $\mathsf{S}_{Q_{X,Z}}$ on β when it comes to satisfying Asm. 12.

Proposition 4. *Let Asm. 17 be satisfied. Let Asm. 12 be satisfied for all $0 \leq \beta \leq \bar{\beta} < +\infty$, where $\mathsf{S}_{Q_{X,Z}} \equiv \mathsf{S}_{Q_{X,Z}}(\beta)$ is bounded in $\mathcal{S}$-norm by $\bar{B}$ for all $\beta \in [0, \bar{\beta}]$. Then, we have that*

$$\gamma_{X,Z} \geq \frac{1}{2} \left[\frac{(\bar{B}^2 C^{2\beta+1} + c^2)\alpha(2\beta+1) - 1}{\bar{B}^2 C^{2\beta+1}\alpha(2\beta+1)} \right].$$

Proof. Write

$$\|\mathbf{M}_{Z|X}\|^2_{\mathrm{HS}(\mathbf{L}^2(Q_Z),\mathbf{L}^2(Q_X))} = \sum_{i \in 1}^\infty \sigma_i^2 = \sum_{i=1}^\infty \mu_i^{2\beta+1} \langle \mathsf{S}_{Q_{X,Z}}, \mu_i^{1/2} e_i \rangle^2_{\mathcal{S}}$$

$$= \|\mathsf{S}_{Q_{X,Z}}\|^2_{\mathcal{S}} \sum_{i=1}^\infty \mu_i^{2\beta+1} \langle \mathsf{S}_{Q_{X,Z}}/\|\mathsf{S}_{Q_{X,Z}}\|_{\mathcal{S}}, \mu_i^{1/2} e_i \rangle^2_{\mathcal{S}}$$

$$\leq \bar{B}^2 \sum_{i=1}^\infty \mu_i^{2(\beta+1/2)}. \tag{83}$$

The right-hand side is finite, for all $\beta \geq 0$, as the $(\mu_i)_{i=1}^\infty$ sequence is associated to a trace class operator. Next, using that $\mu_i^{\beta+1/2} \leq C^{2\beta+1} i^{-(\beta+1/2)\alpha}$, we use Lem. 3 to upper bound (83) via

$$\sum_{i=1}^\infty \mu_i^{2(\beta+1/2)} \leq \frac{C^{2\beta+1}(2\beta+1)\alpha}{(2\beta+1)\alpha - 1} = \frac{C^{2\beta+1}}{1 - (2\beta+1)^{-1}\alpha^{-1}}.$$

On the other hand, using Definition 8 and Lem. 3, the Hilbert-Schmidt norm is lower bounded via

$$\|\mathbf{M}_{Z|X}\|^2_{\mathrm{HS}(\mathbf{L}^2(Q_Z), \mathbf{L}^2(Q_X))} \geq \frac{c^2}{2\gamma_{X,Z} - 1}.$$

Combining both bounds, we have

$$\frac{c^2}{2\gamma_{X,Z} - 1} \leq \frac{\bar{B}^2 C^{2\beta+1}}{1 - (2\beta+1)^{-1}\alpha^{-1}}.$$

Inverting the bound gives the condition

$$\begin{aligned}
\gamma_{X,Z} &\geq \frac{1}{2}\left[\frac{c^2}{\bar{B}^2 C^{2\beta+1}}\left(1 - \frac{1}{\alpha(2\beta+1)}\right) + 1\right] \\
&= \frac{1}{2}\left[\frac{(\bar{B}^2 C^{2\beta+1} + c^2)\alpha(2\beta+1) - 1}{\bar{B}^2 C^{2\beta+1}\alpha(2\beta+1)}\right],
\end{aligned}$$

the result as desired. $\square$

From Prop. 4, we consider the case in which $\alpha \to \infty$ (the data is finite-rank under independence), and derive the singular decay condition

$$\gamma_{X,Z} \geq \frac{1}{2}\left(\frac{\bar{B}^2 C^{2\beta+1} + c^2}{\bar{B}^2 C^{2\beta+1}}\right) > \frac{1}{2}$$

for $c > 0$. While the relationship is not as direct as in the case of (72), we may still observe some regimes in which a "maximally smooth" target function boils down to an independence assumption. This holds intuitively as well, in the sense that $\mathsf{R} \equiv 1$ holds $(Q_X \otimes Q_Z)$-almost surely if and only if X and Z are independent.

D.2.3. CONTROLLING THE PROMPTING TERM

The term that relates $\hat{\rho}_{Y,Z}$ to $\rho_{Y,Z}$ is simply a measurement of the deviation of a sample mean from its population counterpart, within a Hilbert space. Thus, it is reasonable to assume an $O(1/M)$ scaling on this term. Below, we use the notation (X_i', Z_i') to indicate a sample drawn from $Q_X \otimes Q_Z$, i.e., an unpaired example.

Assumption 18. For constants $\delta \in (0,1]$ and $M \geq 1$, there is an event $\mathcal{E}(\delta, M)$, which is independent of the pre-training data $\{(X_i, Z_i)\}_{i=1}^{N/2}$, $\{(X_i', Z_i')\}_{i=1}^{N/2}$, such that on $\mathcal{E}(\delta, M)$,

$$\left\|\tfrac{1}{M}\sum_{j=1}^{M} W_j - \mathbb{E}_{\rho_{Y,Z}}[W_1]\right\|^2_{\mathbf{L}^2(Q_X)} \leq C_{\mathsf{R},\rho}(Q_X)\,\mathrm{plog}(1/\delta)M^{-1}, \tag{84}$$

where $C_{\mathsf{R},\rho}(Q_X)$ depends only on its arguments and r, and is independent of M and δ. On $(\Omega, \mathcal{F}, \mathbb{P})$, the event $\mathcal{E}(\delta, M)$ occurs with probability at least $1 - \delta/2$.

The scaling shown in Asm. 18 can be satisfied by placing a Bernstein-type condition on the random variable W_1 and applying, for instance, the Pinelis-Sahanenko inequality (Pinelis and Sakhanenko, 1986). Specifically, consider the case in which there are positive constants $\sigma, c > 0$ such that

$$\sum_{j=1}^{M} \mathbb{E}_{\rho_{Y,Z}}\|\tfrac{1}{M}W_j - \tfrac{1}{M}\mathbb{E}_{\rho_{Y,Z}}[W_1]\|^q_{\mathbf{L}^2(Q_X)} \leq \frac{q!}{2}\sigma^2 c^{q-2}$$

for all $q \geq 2$. Then, (84) is satisfied, wherein the scalars σ and c will scale as $1/M$, and have additional constants that depend on r, R, $\rho_{Y,Z}$, and Q_X (but not Q_Z or $Q_{X,Z}$). This generates the constant $C_{\mathsf{R},\rho}(Q_X)$ above.

D.2.4. COMPLETING THE PROOF

Well-Specified Case. Because Prop. 4 yields an inexact relationship between the singular decay exponent $\gamma_{X,Z}$ and the source condition constant β, we maintain the statement of the result in terms of this constant. The following result comes as an immediate consequence of Lem. 12 and Cor. 2.

Theorem 11. *Consider failure probability $\delta \in (0, 1]$. Assume that the conditions of Lem. 12 are satisfied and that N is large enough such that the conditions of Cor. 2 are satisfied, in addition to Asm. 18. Define*

$$K_{\max} := 1 + (4\kappa_{\max}^2 + \kappa_{\max})^2.$$

Then, with probability at least $1 - \delta$, it holds that

$$\|\hat{\eta}_\rho - \eta_\rho\|_{\mathbf{L}^2(Q_X)}^2 \lesssim \mathrm{plog}(1/\delta) \left[K_{\max}^{\frac{\beta+2}{\beta+1}} N^{-\frac{\beta}{\beta+1}} + C_{\mathsf{R},\rho}(Q_X) M^{-1} \right] + \chi^2(\rho_Z \| Q_Z),$$

where $C_{\mathsf{R},\rho}(Q_X)$ depends only on its arguments and r, and not M or δ.

The constant $C_{\mathsf{R},\rho}(Q_X)$ appears directly from Asm. 18.

Mis-Specified Case. As mentioned in Appx. B.4, the mis-specified case ($\mathsf{R} \notin \mathcal{S}$) for Radon-Nikodym derivative estimation problems is less understood than the mis-specified case for real-valued and vector-valued nonparametric regression. We intend here to highlight the overall decomposition of error, for which such results could be plugged in as well.

D.3. Distribution Shift

The results of the previous two subsections provided bounds in high probability on the term $\|\hat{\eta}_\rho - \eta_\rho\|_{\mathbf{L}^2(Q_X)}^2$. Returning to the original error decomposition of (12), we would like to relate this to a similar bound on $\|\hat{\eta}_\rho - \eta_\rho\|_{\mathbf{L}^2(P_X)}^2$. We collect two general techniques for performing this change of measure, which lead to either a multiplicative or additive error depending on the assumptions the user is willing to make.

Lemma 14 (Distribution Shift). *Assume that P_X and Q_X have densities p_X and q_X with respect to a common dominating measure ν_X on the measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, and define the* total variation *metric*

$$\mathrm{TV}(P_X, Q_X) := \int_{\mathcal{X}} |p_X(\boldsymbol{x}) - q_X(\boldsymbol{x})| \, d\nu_X(\boldsymbol{x}).$$

Then, for any $\eta : \mathcal{X} \to \mathbb{R}$ such that $[\eta]_X \in \mathbf{L}^2(P_X) \cap \mathbf{L}^2(Q_X)$ (see Appx. B.1), the following holds.

- *If the essential supremum $\|\eta\|_\infty := \inf \left\{ \sup_{A \in \mathcal{B}(\mathcal{X})} \sup_{\boldsymbol{x} \in A} |\eta(\boldsymbol{x})| : \nu_X(A^c) = 0 \right\}$ is finite, then we have the additive relation*

$$\|\eta\|_{\mathbf{L}^2(P_X)}^2 \leq \|\eta\|_{\mathbf{L}^2(Q_X)}^2 + \|\eta\|_\infty^2 \, \mathrm{TV}(P_X, Q_X). \tag{85}$$

- *If $Q_X \ll P_X$, and $\frac{dQ_X}{dP_X}(\boldsymbol{x}, \boldsymbol{z}) \leq B_{P,Q}$ for P_X-almost all $\boldsymbol{x} \in \mathcal{X}$, then we have the multiplicative relation*

$$\|\eta\|_{\mathbf{L}^2(P_X)}^2 \leq B_{P,Q} \|\eta\|_{\mathbf{L}^2(Q_X)}^2. \tag{86}$$

Proof. In the case of (85), we apply Hölder's inequality to achieve

$$\|\eta\|_{\mathbf{L}^2(P_X)}^2 = \mathbb{E}_{P_X}\left[\eta^2(X)\right] = \mathbb{E}_{Q_X}\left[\eta^2(X)\right] + \int_{\mathcal{X}} \eta^2(\boldsymbol{x}) \left(p_X(\boldsymbol{x}) - q_X(\boldsymbol{x})\right) d\nu_X(\boldsymbol{x})$$

$$\leq \|\eta\|_{\mathbf{L}^2(Q_X)}^2 + \|\eta\|_\infty^2 \int_{\mathcal{X}} |p_X(\boldsymbol{x}) - q_X(\boldsymbol{x})| \, d\nu_X(\boldsymbol{x})$$

$$= \|\eta\|_{\mathbf{L}^2(Q_X)}^2 + \|\eta\|_\infty^2 \, \mathrm{TV}(P_X, Q_X),$$

which proves the first claim. For (86), on the other hand, write

$$\|\eta\|_{\mathbf{L}^2(P_X)}^2 = \mathbb{E}_{P_X}\left[\eta^2(X)\right] = \mathbb{E}_{Q_X}\left[\eta^2(X)\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(X)\right] \leq B_{P,Q}\|\eta\|_{\mathbf{L}^2(Q_X)}^2,$$

proving the second claim and completing the proof. $\qquad\square$

From Lem. 14 and the boundedness assumption $|r(\cdot)| \leq B_r$, we alter (12) slightly to read

$$\|\eta_\star - \hat{\eta}_\rho\|_{\mathbf{L}^2(P_X)}^2 \leq 2\|\eta_\star - \eta_\rho\|_{\mathbf{L}^2(P_X)}^2 + 2\|\eta_\rho - \hat{\eta}_\rho\|_{\mathbf{L}^2(Q_X)}^2 + 4B_r^2\,\mathrm{TV}(P_X, Q_X), \tag{87}$$

and plug in the previous bounds on the $\|\eta_\rho - \hat{\eta}_\rho\|_{\mathbf{L}^2(Q_X)}^2$ term for an overall result.

D.4. From Regression to Classification

Throughout this appendix, we evaluated the quality of an estimated map $\hat{\eta}_\rho : \mathcal{X} \to \mathbb{R}$ via its $\mathbf{L}^2(Q_X)$ distance to some target predictor η_ρ. This goal was based on the error decomposition (87), which feeds into ultimate upper bound for $\|\hat{\eta}_\rho - \eta_\star\|_{\mathbf{L}^2(P_X)}^2$, where each term was controlled using the techniques of Appx. C, Appx. D.1, and Appx. D.2. In the case that $r : \mathcal{Y} \to \mathbb{R}$ represents a classification or structured prediction problem (e.g. $r(\boldsymbol{y}) = \mathbb{1}\{\boldsymbol{y} = c\}$ for class $c \in \mathcal{Y}$), it is of clear interest whether the control of mean squared error translates to risk guarantees for classification error. Establishing these guarantees, using the notion of a *structure encoding loss function (SELF)* described in Bach (2024, Section 13.2), is the subject of this section.

Assume that $\mathcal{Y}$ is discrete, or that $|\mathcal{Y}| < \infty$. We consider a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and a regular conditional distribution $P_{Y|X}(\cdot|\boldsymbol{x})$ (see Definition 4), under which $\ell(\cdot, \boldsymbol{y})$ is integrable for all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$. The corresponding risk of any map $h : \mathcal{X} \to \mathcal{Y}$ will be denoted

$$\mathcal{R}(h) = \mathbb{E}_{P_{X,Y}}\left[\ell(Y, h(X))\right]. \tag{88}$$

There are a number of assumptions that mark the SELF framework.

Assumption 19 (SELF Loss for Structured Prediction). Consider the existence of a Hilbert space $\mathcal{F}$, and two mappings $\chi : \mathcal{Y} \to \mathcal{F}$ and $\xi : \mathcal{Y} \to \mathcal{F}$ which act as embeddings of objects in $\mathcal{Y}$. Then, assume that ℓ satisfies the equality

$$\ell(\boldsymbol{y}, \boldsymbol{y}') = \langle \chi(\boldsymbol{y}), \xi(\boldsymbol{y}')\rangle_{\mathcal{F}}.$$

As of yet, no assumptions (such as being an RKHS) have been placed on $\mathcal{F}$. Under Asm. 19, the Bayes optimal predictor (with respect to (88), and not mean squared error) is given by

$$h_\star(\boldsymbol{x}) \in \arg\min_{\boldsymbol{y}' \in \mathcal{Y}} \sum_{\boldsymbol{y} \in \mathcal{Y}} \ell(\boldsymbol{y}, \boldsymbol{y}')P_{Y|X}(\boldsymbol{y}|\boldsymbol{x}),$$

where ties can be broken arbitrarily. In other words, $h_\star \in \arg\min_h \mathcal{R}(h)$. Additionally, because $\mathcal{Y}$ is finite, we may take the expectation

$$\sum_{\boldsymbol{y} \in \mathcal{Y}} \ell(\boldsymbol{y}, \boldsymbol{y}')P_{Y|X}(\boldsymbol{y}|\boldsymbol{x}) = \sum_{\mathcal{Y}} \langle \chi(\boldsymbol{y}), \xi(\boldsymbol{y}')\rangle_{\mathcal{F}} P_{Y|X}(\boldsymbol{y}|\boldsymbol{x})$$

$$= \langle \mathbb{E}_{P_{X,Y}}\left[\chi(Y)|X\right](\boldsymbol{x}), \xi(\boldsymbol{y}')\rangle_{\mathcal{F}},$$

which is only based on finite sums of vectors in $\mathcal{F}$. Next, we define the notation of a surrogate loss. To construct a predictor (e.g. classifier), we consider a function $s : \mathcal{X} \to \mathcal{F}$ called the *score function* and a map $\mathrm{dec} : \mathcal{F} \to \mathcal{Y}$ known as a *decoder*. We will then define an integrable surrogate loss $L : \mathcal{Y} \times \mathcal{F} \to \mathbb{R}$, for which we can define the risk

$$\mathcal{R}^L(s) = \mathbb{E}_{P_{X,Y}}\left[L(Y, s(X))\right]. \tag{89}$$

We can then define the Bayes surrogate risk[5] as

$$\mathcal{R}_\star^L = \mathbb{E}_{P_X} \left[\inf_{h \in \mathcal{F}} \mathbb{E}_{P_{X,Y}} \left[L(Y, h) | X \right] \right].$$

The relationship between the surrogate risk (89) and the true risk (88) for squared surrogates is given in the following result.

Proposition 5. *(Bach, 2024, Section 13.4.2) Consider the surrogate loss and decoder given by*

$$L(\boldsymbol{y}, s(\boldsymbol{x})) := \|\xi(\boldsymbol{y}) - s(\boldsymbol{x})\|_{\mathcal{F}}^2 \text{ and } \mathrm{dec}(h) \in \underset{\boldsymbol{y} \in \mathcal{Y}}{\arg\min} \langle \chi(\boldsymbol{y}), h \rangle_{\mathcal{F}}.$$

Then, for any score function $s : \mathcal{X} \to \mathcal{F}$, it holds that

$$\mathcal{R}(\mathrm{dec} \circ s) - \mathcal{R}(h_\star) \le 2 \sup_{\boldsymbol{y} \in \mathcal{Y}} \|\chi(\boldsymbol{y})\|_{\mathcal{F}} \cdot \sqrt{\mathcal{R}^L(s) - \mathcal{R}_\star^L}.$$

We stated Prop. 5 generally; we now map it to classification, the prototypical task associated with zero-shot prediction. Let $\mathcal{Y} = \{1, \dots, C\}$, where C denotes the number of classes (in contrast to the absolute constants in Thm. 2 and Thm. 3). Then, we have that $\chi(\boldsymbol{y})$ is the one-hot encoding in $\mathbb{R}^C$, whereas $\xi(\boldsymbol{y})$ is the complement, that is, $\xi_j(\boldsymbol{y}) = 1 - \chi_j(\boldsymbol{y})$ for $c = 1, \dots, C$. Thus, their inner product generates the 0-1 loss

$$\ell(\boldsymbol{y}, \boldsymbol{y}') = \mathbb{1}\{\boldsymbol{y} \ne \boldsymbol{y}'\} = \langle \chi(\boldsymbol{y}), \xi(\boldsymbol{y}') \rangle_{\mathbb{R}^C}.$$

Then, we immediately have that $\sup_{\boldsymbol{y} \in \mathcal{Y}} \|\chi(\boldsymbol{y})\|_{\mathcal{F}} = 1$. It remains to determine the score function $s : \mathcal{X} \to \mathbb{R}^C$. Note that we used a function r to define (3) and (4); we will now use C such functions $r^{(1)}, \dots, r^{(C)}$ each defined by

$$r^{(c)}(\boldsymbol{y}) = \xi_j(\boldsymbol{y}) = \mathbb{1}\{\boldsymbol{y} = j\} \tag{90}$$

which in turn gives us the individual mean squared error minimizers

$$\eta_\star^{(c)}(\boldsymbol{x}) = \mathbb{E}_{P_{Y,X}} \left[r^{(c)}(Y) | X \right] (\boldsymbol{x}) = \mathbb{P}_{P_{Y,X}} \left[Y = j | X \right] (\boldsymbol{x}).$$

Finally, we may use any of the estimation strategies developed in Appx. D.1 or Appx. D.2 to produce estimators $\hat{\eta}_\rho^{(1)}, \dots, \hat{\eta}_\rho^{(C)}$ (i.e. the predicted probability per class) to give the score function

$$s(\boldsymbol{x}) := \left(\hat{\eta}_\rho^{(1)}(\boldsymbol{x}), \dots, \hat{\eta}_\rho^{(C)}(\boldsymbol{x}) \right) \in \mathbb{R}^C. \tag{91}$$

Each $\hat{\eta}_\rho^{(c)}$ is then associated to the conditional mean given by the prompt distribution, which we denote $g_\rho^{(c)}$. As a final step, we use the classical relationship between mean squared prediction error and mean squared integrated error, as seen below.

Corollary 3. *For the score function given in (91) and decoder given in Prop. 5, it holds that*

$$\mathcal{R}(\mathrm{dec} \circ s) - \mathcal{R}(h_\star) \le 2 \sqrt{\sum_{j=1}^C \|\hat{\eta}_\rho^{(c)} - \eta_\star^{(c)}\|_{\mathbf{L}^2(P_X)}^2}.$$

Proof. Given Prop. 5, we need only show that

$$\mathcal{R}^L(s) - \mathcal{R}_\star^L = \sum_{j=1}^C \|\hat{\eta}_\rho^{(c)} - \eta_\star^{(c)}\|_{\mathbf{L}^2(P_X)}^2. \tag{92}$$

First, note that for the score function s given in (91), it holds by (90) that

$$L(\boldsymbol{y}, s(\boldsymbol{x})) := \|\xi(\boldsymbol{y}) - s(\boldsymbol{x})\|_{\mathbb{R}^C}^2 = \sum_{j=1}^C (r^{(c)}(\boldsymbol{y}) - \hat{\eta}_\rho^{(c)}(\boldsymbol{x}))^2,$$

---

[5]We assume the map $\boldsymbol{x} \mapsto \inf_{h \in \mathcal{F}} \mathbb{E}_{P_{X,Y}} \left[L(Y, h) | X \right] (\boldsymbol{x})$ to be measurable as a technical consideration.
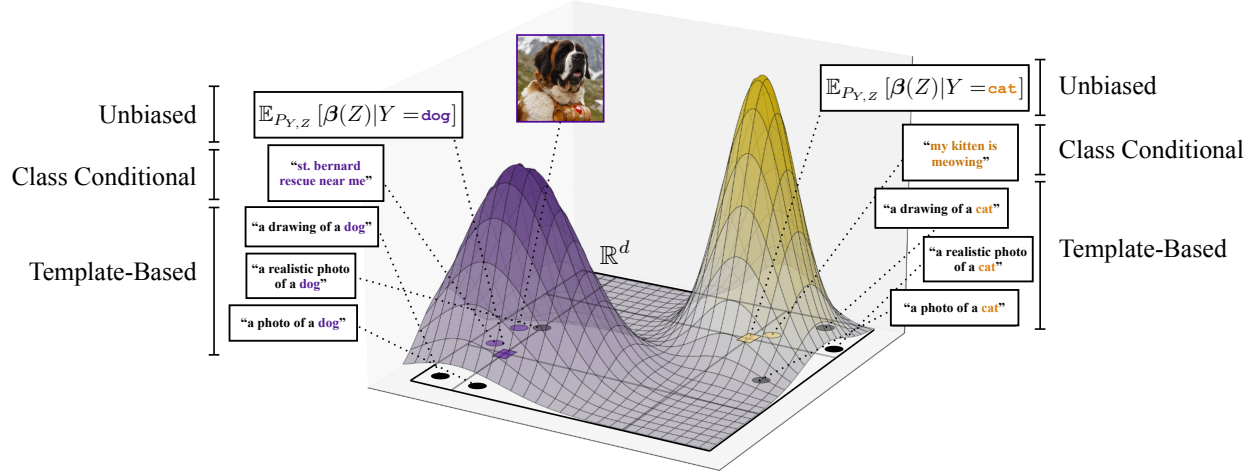
Figure 5. **Illustration of Prompting Strategies.** A hypothetical distribution of embeddings $\beta(Z)$ parametrized by two classes ("cat" and "dog"). Three prompting strategies (template-based, class-conditional, and unbiased) are shown with example text and resulting embeddings in $\mathbb{R}^d$. Colors represent the probability of each class given the embedding.

and after taking the expectation over $P_{X,Y}$,

$$\mathcal{R}^L(s) = \mathbb{E}_{P_{X,Y}}\left[L(Y, s(X))\right] = \sum_{j=1}^{C} \mathbb{E}_{P_{X,Y}}\left[(r^{(c)}(Y) - \hat{\eta}_\rho^{(c)}(X))^2\right].$$

Then, by the bias-variance decomposition for each $c = 1, \ldots, C$, it holds that

$$\underbrace{\sum_{j=1}^{C} \mathbb{E}_{P_{X,Y}}\left[(r^{(c)}(Y) - \hat{\eta}_\rho^{(c)}(X))^2\right]}_{\mathcal{R}^L(s)} = \sum_{j=1}^{C}\|\hat{\eta}_\rho^{(c)} - \eta_\star^{(c)}\|_{\mathbf{L}^2(P_X)}^2 + \underbrace{\sum_{j=1}^{C} \mathbb{E}_{P_{X,Y}}\left[(r^{(c)}(Y) - \eta_\star^{(c)}(X))^2\right]}_{\mathcal{R}_\star^L}.$$

Rearranging terms gives (92) and completes the proof. □

In particular, when applying the bound above to results of Thm. 1, Thm. 2, and Thm. 3, we derive a bound of the form

$$\mathcal{R}(\text{dec} \circ s) - \mathcal{R}(h_\star) \lesssim \sqrt{C\mathbb{E}_{P_Z}\left[I(X;Y|Z)\right] + \sum_{j=1}^{C}\|g_\rho^{(c)} - g_{P_{Y,Z}}^{(c)}\|_{\mathbf{L}^2(P_Z)}^2} + C\,\mathrm{TV}(P_X, Q_X)$$

$$+ \begin{cases} \sqrt{C}\,\mathrm{plog}\,(C/\delta)\left(N^{-\frac{q(t)}{2(q(t)+1)}} + M^{-\frac{2\omega_\rho-1}{4\omega_\rho+2}}\right) & \text{(conditional mean)} \\ \sqrt{C}\,\mathrm{plog}\,(C/\delta)\left(N^{-\frac{\beta}{2(\beta+1)}} + M^{-1/2}\right) + \sqrt{D_{\chi^2}(\rho_Z\|Q_Z)} & \text{(information density)} \end{cases},$$

which holds with probability at least $1-\delta$. While generalization bounds for classification and structured prediction can have sharper dependences on the number of examples and number of classes for supervised learning (e.g., via the techniques of Cabannnes et al. (2021) and references therein), the conversion from regression to classification is a remarkably general way to account for the residual dependence, prompt bias, and multiple stages of estimation that mark our problem.

D.5. Prompting Strategies

We have stated upper bounds on the statistical error in this section that depend on the size of the pre-training set N and the number of prompts M. To state them more precisely, however, we must also specify the sampling schemes that lead to these examples/prompts. Sampling of the pre-training data falls into fixed and well-understood categories, boiling down to whether only paired examples or a combination of paired and unpaired examples are available. We describe

these as part of the background (Appx. B.4), alongside the method to which they apply. However, the interpretation of prompting (the empirical procedure used in (1)) formally as a sampling scheme from a probability measure $\rho_{Y,Z}$ is itself a contribution of this paper. In the results of Appx. D.1 and Appx. D.2, we considered simple random sampling $(Y_1, Z_1), \ldots, (Y_M, Z_M) \sim \rho_{Y,Z}$ i.i.d. to provide examples of scenarios in which Asm. 14 and Asm. 18 can be satisfied. However, multiple practical and idealized strategies exist for prompting (such as the ones explored in Sec. 4). Below, we represent them in our framework below, as ways to define $\rho_{Y,Z}$ and approximate it with $\hat{\rho}_{Y,Z}$.

- **Template-Based:** This technique reflects the earlier iterations of representing labels in natural language. Examples include "photo of a ___", "realistic photo of a ___", "drawing of a ___", etc. Notice that the prompt templates have no relationship with the class label. One way this can be understood is by representing the caption via the structural equation $Z = f(Y, U)$, where U represents the text of the caption with the label left blank (drawn according to a probability measure ρ_U), and f represents the action of inserting the natural language label. Then, we have that under the template-based prompting distribution, $U \perp\!\!\!\perp Y$. This does not imply that $Z \perp\!\!\!\perp Y$, but instead that the dependence is governed fully by the function f. To sample, a fixed number of m examples $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ are drawn directly from ρ_U. We then use the empirical measure $\hat{\rho}_{Y,Z}(\boldsymbol{y}, \boldsymbol{z}) = \frac{1}{m} \sum_{k=1}^{m} \rho_Y(\boldsymbol{y}) \mathbb{1}\{f(\boldsymbol{y}, \boldsymbol{u}_k) = \boldsymbol{z}\}$, where ρ_Y is fixed as the uniform distribution on the discrete set $\mathcal{Y}$. Here, $M = m\,|\mathcal{Y}|$.

- **Class Conditional:** This technique reflects the modern LLM-based techniques, such as CuPL (Pratt et al., 2023). We parameterize the joint distribution using the conditional distributions $\rho_{Y,Z} = \sum_{\boldsymbol{y} \in \mathcal{Y}} \rho_{Z|Y=\boldsymbol{y}} \cdot \rho_Y(\boldsymbol{y})$ for each class $\boldsymbol{y} \in \mathcal{Y}$. Sampling from each $\rho_{Z|Y=\boldsymbol{y}}$ occurs by meta-prompting the LLM (such as the one we use in Appx. F), which generates samples $\boldsymbol{z}_1^{\boldsymbol{y}}, \ldots, \boldsymbol{z}_1^{\boldsymbol{y}}$ and empirical measures $\hat{\rho}_{Z|Y=\boldsymbol{y}} = \frac{1}{m} \sum_{k=1}^{m} \delta_{\boldsymbol{z}_k^{\boldsymbol{y}}}$. Our final approximation is $\hat{\rho}_{Y,Z} = \sum_{\boldsymbol{y} \in \mathcal{Y}} \hat{\rho}_{Z|Y=\boldsymbol{y}} \cdot \rho_Y(\boldsymbol{y})$, with $M = m\,|\mathcal{Y}|$.

- **Unbiased:** This techniques reflects the setting of Fig. 3, where the user may drawn samples from a joint distribution $P_{X,Y,Z}$, where the marginal $P_{X,Y}$ is in fact the data on which the zero-shot classifier will be evaluated. Then, the prompt distribution can be constructed, as we do, by drawing samples $(\boldsymbol{y}_1, \boldsymbol{z}_1), \ldots, (\boldsymbol{y}_M, \boldsymbol{z}_M)$ directly from $P_{Y,Z}$, and defining $\hat{\rho}_{Y,Z} = \frac{1}{M} \sum_{j=1}^{M} \delta_{(\boldsymbol{y}_j, \boldsymbol{z}_j))}$. We call this "unbiased", because the prompt bias term in Thm. 1 is zero for this example. It is worth pointing out that even if $P_{Y|Z=\boldsymbol{z}}$ can be matched by the prompt distribution, the distribution mismatch term from Thm. 3 will be zero if and only if $\rho_Z = Q_Z$ (or the prompt captions match the pre-training captions in distribution). In this sense, $P_{Y,Z}$ may not be the ideal prompting distribution, but instead, $P_{Y|Z}Q_Z$.

E. Self-Supervised Objectives and Cross Covariance Operators

In Sec. 3, we considered specific instances of both the conditional mean and information density approaches based on nonparametric regression in reproducing kernel Hilbert space (RKHS). This reflected the statistical goals of Thm. 2 and Thm. 3. In this appendix, we aim to draw relationships with other approaches based on optimizing self-supervised learning (SSL) objectives, in order to align with practice. In particular, we focus on the relationship between such objectives and the mean square contingency $I(X; Z)$ introduced in Sec. 2. To do so, we make explicit the intuition that SSL objectives (such as CLIP and VICReg) are implicit forms of dependence maximization between the representations $\boldsymbol{\alpha}(X)$ and $\boldsymbol{\beta}(Z)$. Some of the arguments below have previously appeared in the literature—we do not claim originality for them, but instead aim to consolidate them together in a single vignette.

When it comes to specific SSL objectives, we describe here their properties as functions acting on a batch of encoded data $(\boldsymbol{\alpha}(\boldsymbol{x}_1), \boldsymbol{\beta}(\boldsymbol{z}_1)), \ldots, (\boldsymbol{\alpha}(\boldsymbol{x}_n), \beta(\boldsymbol{z}_n))$. This abstract description is agnostic to the function class used for the encoder. Reproducing kernel Hilbert space theory has been frequently used, in the recent literature, to define the function classes involved in contrastive and non-contrastive self-supervised foundation modeling (Li et al., 2021; Balestriero and LeCun, 2022; Kiani et al., 2022; Johnson et al., 2023; Tan et al., 2024). We also mention that the precise characterization of the function classes of various deep neural networks is an active area of research (Schmidt-Hieber, 2020; Scetbon and Harchaoui, 2020; Parhi and Nowak, 2021; Wu and Long, 2022; Bartolucci et al., 2023; Unser, 2023; Siegel and Xu, 2023; Shwartz-Ziv et al., 2023; DeVore et al., 2025). However, these exciting yet still burgeoning theories of deep neural networks have not yet reached a maturity level comparable to the one of RKHS theory (Wahba, 1990; Cucker and Zhou, 2007; Christmann and Steinwart, 2008; Bach, 2024) needed for the theoretical analysis we develop in this paper. For more practical details on self-supervised learning, we point the reader to the recent survey (Balestriero et al., 2023).

Covariance Operators. To relate our theory (which centers around the mean square contingency measure of dependence) to SSL objectives, we first draw the relationship to covariance operators of $Q_{X,Z}$ on particular function spaces. Let $\mathcal{H}$ be an RKHS of real-valued functions $\mathcal{X}$ and $\mathcal{G}$ be an RKHS of real-valued functions on $\mathcal{Z}$. Then, define the cross-covariance operator $\mathbf{C}_{XZ} : \mathcal{G} \to \mathcal{H}$ by

$$\langle h, \mathbf{C}_{XZ}g \rangle_{\mathcal{H}} = \mathbb{C}\mathrm{ov}_{Q_{X,Z}}(h(X), g(Z)),$$

and the analogously defined auto-covariance operators $\mathbf{C}_{XX} : \mathcal{H} \to \mathcal{H}$ and $\mathbf{C}_{ZZ} : \mathcal{G} \to \mathcal{G}$. When $\mathbf{C}_{XX}$ and $\mathbf{C}_{ZZ}$ are compact, we define the powers $\mathbf{C}_{XX}^{1/2}$ and $\mathbf{C}_{ZZ}^{1/2}$ in the sense of (39). It then holds by Baker (1973, Theorem 1) that there exists a unique bounded linear operator $\mathbf{V}_{XZ} : \mathcal{G} \to \mathcal{H}$, so that

$$\mathbf{C}_{XZ} = \mathbf{C}_{XX}^{1/2} \mathbf{V}_{XZ} \mathbf{C}_{ZZ}^{1/2}. \tag{93}$$

The operator $\mathbf{V}_{XZ}$ is called the *normalized cross-covariance operator*, or NOCCO for short (Fukumizu et al., 2005). As an abuse of notation, the NOCCO (93) is sometimes communicated as $\mathbf{V}_{XZ} = \mathbf{C}_{XX}^{-1/2} \mathbf{C}_{XZ} \mathbf{C}_{ZZ}^{-1/2}$, though it is uniquely defined without necessarily constructing the square-root inverses. To rigorously use the formula $\mathbf{C}_{XX}^{-1/2} \mathbf{C}_{XZ} \mathbf{C}_{ZZ}^{-1/2}$ with a well-defined adjoint, we must make assumptions on the closure of the range of $\mathbf{C}_{XZ}$ and $\mathbf{C}_{ZX}$ being contained within the closure of the range of $\mathbf{C}_{XX}$ and $\mathbf{C}_{ZZ}$, respectively. The Hilbert-Schmidt norm of the population NOCCO, when finite, is equal to the mean square contingency

$$\|\mathbf{V}_{X,Z}\|_{\mathrm{HS}(\mathcal{G},\mathcal{H})}^2 = I(X; Z), \tag{94}$$

as shown in Fukumizu et al. (2007b, Theorem 4). The relation (94) requires a few additional technical conditions, such as $(\mathcal{H} \otimes \mathcal{G}) + \mathbb{R}$ being dense in $\mathbf{L}^2(Q_X \otimes Q_Z)$ and $Q_{X,Z}$ having joint and marginal densities[6].

Variational Characterization of the Hilbert-Schmidt Norm. This operator is an essential component of the kernel canonical correlations analysis (CCA) problem, which (with (94)) will be the common bridge that ties together SSL and the mean square contingency. From the nonparametric CCA perspective, the singular values $(\sigma_i)_{i=1}^{\infty}$ refer precisely to the canonical correlations and the singular functions $((\alpha_i, \beta_i))_{i=1}^{\infty}$ refer to the canonical variates (Lancaster, 1958; Buja, 1990; Michaeli et al., 2016). Returning to (94), this operator is estimated with a regularization scheme, i.e.

$$\widehat{\mathbf{V}}_{X,Z} := (\widehat{\mathbf{C}}_{XX} + \lambda \mathbf{I})^{-1/2} \widehat{\mathbf{C}}_{XZ} (\widehat{\mathbf{C}}_{ZZ} + \lambda \mathbf{I})^{-1/2},$$

where $\widehat{\mathbf{C}}_{XX}$, $\widehat{\mathbf{C}}_{XZ}$, and $\widehat{\mathbf{C}}_{ZZ}$ are the standard empirical covariance estimates (see Appx. B.4) and $\lambda > 0$ is a regularization parameter. Then, one solves the empirical CCA problem

$$\max_{\substack{h_1,\ldots,h_d \in \mathcal{H} \text{ o.n.b} \\ g_1,\ldots,g_d \in \mathcal{G} \text{ o.n.b}}} \sum_{i=1}^{d} \langle h_i, \widehat{\mathbf{V}}_{X,Z} g_i \rangle_{\mathcal{H}}. \tag{95}$$

where o.n.b denotes an orthonormal basis. Setting aside matters of estimation, we consider how the norm quantity $\|\mathbf{V}_{X,Z}\|_{\mathrm{HS}(\mathcal{G},\mathcal{H})}^2$ relates to the actual encoders returned by the CCA problem (95) (assuming that $\widehat{\mathbf{V}}_{X,Z} \approx \mathbf{V}_{X,Z}$). Let $(s_i)_{i=1}^{\infty}$ be ordered singular values of the Hilbert-Schmidt operator $\mathbf{V}_{X,Z}$. In this case, denoting $h_1, \ldots, h_d \in \mathcal{H}$ and

---

[6]Recall that our definition of $I(X; Z)$ does not include the square root that is usually used in the definition of mean square contingency.

$g_1, \ldots, g_d \in \mathcal{G}$ the orthonormal bases of $\mathcal{H}$ and $\mathcal{G}$ resp. maximizing the criterion ((95)), we have

$$\sum_{i=1}^{d} \langle h_i, \mathbf{V}_{X,Z} g_i \rangle_{\mathcal{H}} = \sum_{i=1}^{d} s_i \leq \sqrt{d \sum_{i=1}^{d} s_i^2} \tag{96}$$

$$= \sqrt{d \left(\|\mathbf{V}_{X,Z}\|_{\mathrm{HS}(\mathcal{G},\mathcal{H})}^2 - \sum_{i=d+1}^{\infty} s_i^2 \right)}$$

$$= \sqrt{d \left(I(X;Z) - \sum_{i=d+1}^{\infty} s_i^2 \right)} \tag{97}$$

The two orthonormal bases maximizing the criterion (95) are actually the left and right singular functions of $\mathbf{V}_{XY}$ associated to the leading d singular values (see Thm. 5). The larger the truncation level d, the closer the quantity is to the mean-square contingency, up to the truncation level factor d.

In either the population (96) or empirical (95) problems, the functions are maximizing an objective that is a measure of covariance with a constraint on variance. The constraint on variance is imposed by the norm condition on $h_1, \ldots, h_d \in \mathcal{H}$ and $g_1, \ldots, g_d \in \mathcal{G}$, respectively); see (Fukumizu et al., 2007a). This norm condition is relaxed into a penalization term in popular SSL objectives.

Indeed, several SSL objectives can be written in an analogous *variance-regularized covariance* form. This may offer one intuitive viewpoint as to why estimators based on these objectives might exhibit similar statistical properties to those analyzed in Sec. 3. We first describe a format for these variance-regularized covariance objectives and show that a number of popular SSL objectives can be expressed in this form.

Variance-Regularized Covariance Objectives. Recall that $\boldsymbol{\alpha} : \mathcal{X} \to \mathbb{R}^d$ and $\boldsymbol{\beta} : \mathcal{Z} \to \mathbb{R}^d$ denote encoders for $\mathcal{X}$-valued and $\mathcal{Z}$-valued objects (often images and text, respectively). We denote the standard Euclidean inner product by $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \sum_{j=1}^{d} u_j v_j$ in $\mathbb{R}^d$. In either case, we consider a batch of data points $(\boldsymbol{x}_1, \boldsymbol{z}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{z}_n)$ which are thought to be n independent and identically distributed realizations of (X, Z) from the probability distribution $Q_{X,Z}$ over $\mathcal{X} \times \mathcal{Z}$. Let us then define the design matrices induced by the embeddings, written as

$$\mathbf{A} := \begin{bmatrix} - \boldsymbol{\alpha}(\boldsymbol{x}_1) - \\ \vdots \\ - \boldsymbol{\alpha}(\boldsymbol{x}_n) - \end{bmatrix} \in \mathbb{R}^{n \times d} \text{ and } \mathbf{B} := \begin{bmatrix} - \boldsymbol{\beta}(\boldsymbol{z}_1) - \\ \vdots \\ - \boldsymbol{\beta}(\boldsymbol{z}_n) - \end{bmatrix} \in \mathbb{R}^{n \times d}.$$

Let $\mathbf{J} := \mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}^\top \in \mathbb{R}^{n \times n}$ be the centering matrix and construct the empirical auto-covariance and cross-covariance matrices

$$\hat{\boldsymbol{\Sigma}}_{AA} := (\mathbf{J}\mathbf{A})^\top (\mathbf{J}\mathbf{A}), \quad \hat{\boldsymbol{\Sigma}}_{BB} := (\mathbf{J}\mathbf{B})^\top (\mathbf{J}\mathbf{B}), \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}_{AB} := (\mathbf{J}\mathbf{A})^\top (\mathbf{J}\mathbf{B}).$$

We aim to write the upcoming objectives in the form

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = -\mathbf{Tr}\left(\hat{\boldsymbol{\Sigma}}_{AB} \right) + \kappa \|\bar{\boldsymbol{\Sigma}}_{AB}\|_{\mathrm{F}}^2 + V(\boldsymbol{\alpha}, \boldsymbol{\beta}),$$

for hyperparameter $\kappa \geq 0$, matrix $\bar{\boldsymbol{\Sigma}}_{AB}$ (which is $\hat{\boldsymbol{\Sigma}}_{AB}$ with its diagonal components set to zero), and variance-regularization term $V(\boldsymbol{\alpha}, \boldsymbol{\beta})$. The term $V(\boldsymbol{\alpha}, \boldsymbol{\beta})$ may explicitly include the regularized inverses of $\hat{\boldsymbol{\Sigma}}_{XX}$ and $\hat{\boldsymbol{\Sigma}}_{ZZ}$, or may penalize variance or non-smoothness more implicitly.

Example 1: Multimodal InfoNCE (CLIP). Consider the empirical objective for the contrastive language-image pre-training (CLIP) model (Radford et al., 2021) with batch size n,

$$\hat{\mathcal{L}}_{\mathrm{CLIP}}(\boldsymbol{\alpha},\boldsymbol{\beta}) := -\frac{1}{n}\sum_{i=1}^{n}\langle\boldsymbol{\alpha}(\boldsymbol{x}_i),\boldsymbol{\beta}(\boldsymbol{z}_i)\rangle + \frac{1}{2}\log\sum_{j=1}^{n}e^{\langle\boldsymbol{\alpha}(\boldsymbol{x}_i),\boldsymbol{\beta}(\boldsymbol{z}_j)\rangle} + \frac{1}{2}\log\sum_{j=1}^{n}e^{\langle\boldsymbol{\alpha}(\boldsymbol{x}_j),\boldsymbol{\beta}(\boldsymbol{z}_i)\rangle} + \log n,$$

where the $\log n$ factor is appended to normalize the sums in the logarithmic terms and does not change the minimizer. Following arguments used (e.g. by Li et al. (2021)) for the SimCLR objective—the single-modality counterpart to CLIP—we analyze the logarithmic terms via Tayler expansion. To simplify the analysis, take the large-sample limit to define the population objective

$$\mathcal{L}_{\mathrm{CLIP}}(\boldsymbol{\alpha},\boldsymbol{\beta}) := -\mathbb{E}_P\langle\boldsymbol{\alpha}(X),\boldsymbol{\beta}(Z)\rangle$$
$$+ \frac{1}{2}\mathbb{E}_{P_X}\left[\log\mathbb{E}_P\left[e^{\langle\boldsymbol{\alpha}(X),\boldsymbol{\beta}(Z)\rangle}\Big|X\right]\right] + \frac{1}{2}\mathbb{E}_{P_Z}\left[\log\mathbb{E}_P\left[e^{\langle\boldsymbol{\alpha}(X),\boldsymbol{\beta}(Z)\rangle}\Big|Z\right]\right]. \tag{98}$$

Next, define the quantity $c(\boldsymbol{x}) := \mathbb{E}_{P_{XZ}}\left[\langle\boldsymbol{\alpha}(X),\boldsymbol{\beta}(Z)\rangle|X\right](\boldsymbol{x})$ and apply a second-order Taylor expansion for every $\boldsymbol{x} \in \mathcal{X}$, the approximation

$$e^{\langle\boldsymbol{\alpha}(\boldsymbol{x}),\boldsymbol{\beta}(Z)\rangle} = e^{c(\boldsymbol{x})}e^{\langle\boldsymbol{\alpha}(\boldsymbol{x}),\boldsymbol{\beta}(Z)\rangle-c(\boldsymbol{x})}$$
$$\approx e^{c(\boldsymbol{x})}\left(1 + \langle\boldsymbol{\alpha}(\boldsymbol{x}),\boldsymbol{\beta}(Z)\rangle - c(\boldsymbol{x}) + \frac{1}{2}\left(\langle\boldsymbol{\alpha}(\boldsymbol{x}),\boldsymbol{\beta}(Z)\rangle - c(\boldsymbol{x})\right)^2\right).$$

Plugging this approximation into the first term of (98) yields

$$\log\mathbb{E}_{P_Z}\left[e^{\langle\boldsymbol{\alpha}(X),\boldsymbol{\beta}(Z)\rangle}\Big|X\right](\boldsymbol{x}) \approx c(\boldsymbol{x}) + \log\left(1 + \tfrac{1}{2}\mathbb{V}\mathrm{ar}\left(\langle\boldsymbol{\alpha}(\boldsymbol{x}),\boldsymbol{\beta}(Z)\rangle|X\right)(\boldsymbol{x})\right)$$

Using the Taylor expansion $\log(1 + y) = y + o(y)$ centered at $y = 0$, and evaluate the first-order approximation at $y = \frac{1}{2}\mathbb{V}\mathrm{ar}\left(\langle\boldsymbol{\alpha}(\boldsymbol{x}),\boldsymbol{\beta}(Z)\rangle|X\right)(\boldsymbol{x})$, we finally have that

$$\frac{1}{2}\mathbb{E}_{P_X}\left[\log\mathbb{E}_{P_Z}\left[e^{\langle\boldsymbol{\alpha}(X),\boldsymbol{\beta}(Z)\rangle}\Big|X\right]\right] \approx \frac{1}{2}\mathbb{E}_P\left[\langle\boldsymbol{\alpha}(X),\boldsymbol{\beta}(Z)\rangle\right] + \frac{1}{4}\mathbb{E}_{P_X}\left[\mathbb{V}\mathrm{ar}\left(\langle\boldsymbol{\alpha}(X),\boldsymbol{\beta}(Z)\rangle|X\right)\right].$$

Applying an identical argument to the second term of (98) gives

$$\mathcal{L}_{\mathrm{CLIP}}(\boldsymbol{\alpha},\boldsymbol{\beta}) \approx -\left(\mathbb{E}_P\langle\boldsymbol{\alpha}(X),\boldsymbol{\beta}(Z)\rangle - \langle\mathbb{E}_{P_X}\left[\boldsymbol{\alpha}(X)\right],\mathbb{E}_{P_Z}\left[\boldsymbol{\beta}(Z)\right]\rangle\right)$$
$$+ \tfrac{1}{4}\mathbb{E}_{P_X}\left[\mathbb{V}\mathrm{ar}\left(\langle\boldsymbol{\alpha}(X),\boldsymbol{\beta}(Z)\rangle|X\right)\right] + \tfrac{1}{4}\mathbb{E}_{P_Z}\left[\mathbb{V}\mathrm{ar}\left(\langle\boldsymbol{\alpha}(X),\boldsymbol{\beta}(Z)\rangle|Z\right)\right]$$
$$= -\mathbf{Tr}\left(\mathbb{C}\mathrm{ov}(\boldsymbol{\alpha}(X),\boldsymbol{\beta}(Z))\right) + \tfrac{1}{4}\mathbb{E}_{P_X}\left[\mathbb{V}\mathrm{ar}\left(\langle\boldsymbol{\alpha}(X),\boldsymbol{\beta}(Z)\rangle|X\right)\right] + \tfrac{1}{4}\mathbb{E}_{P_Z}\left[\mathbb{V}\mathrm{ar}\left(\langle\boldsymbol{\alpha}(X),\boldsymbol{\beta}(Z)\rangle|Z\right)\right].$$

which is the desired form for the population. Now, to rewrite the empirical version, we have

$$\hat{\mathcal{L}}_{\mathrm{CLIP}}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \underbrace{-\mathbf{Tr}\left(\hat{\boldsymbol{\Sigma}}_{AB}\right)}_{\text{covariance}} + \underbrace{\frac{1}{4N}\sum_{i=1}^{N}\widehat{\mathbb{V}\mathrm{ar}}_N\left(\langle\boldsymbol{\alpha}(X),\boldsymbol{\beta}(Z)\rangle|X\right)(\boldsymbol{x}_i) + \frac{1}{4N}\sum_{i=1}^{N}\widehat{\mathbb{V}\mathrm{ar}}_N\left(\langle\boldsymbol{\alpha}(X),\boldsymbol{\beta}(Z)\rangle|Z\right)(\boldsymbol{z}_i)}_{\text{variance regularization}},$$

where $\widehat{\mathbb{V}\mathrm{ar}}_N$ denotes the variance with respect to the empirical measure $\frac{1}{N}\sum_{i=1}^{N}\delta_{(\boldsymbol{x}_i,\boldsymbol{z}_i)}$.

Example 2: BarlowTwins. The BarlowTwins objective (Zbontar et al., 2021) has already been interpreted as an instance of kernel canonical correlations analysis (CCA) by previous work (e.g. by Balestriero and LeCun (2022)). This objective is usually defined in terms of the cross-correlation and auto-correlation matrices. To be consistent with other objectives in this section, we handle this by enforcing a constraint on the variance. Let $\iota_S : \mathbb{R}^{d\times d} \to \{0, +\infty\}$ denote the convex analytic indicator function such that $\iota(\bar{\boldsymbol{\Sigma}}) = 0$ if $\bar{\boldsymbol{\Sigma}} \in S$ and equals $+\infty$ otherwise. Let $\mathbf{I}$ be the identity matrix in $\mathbb{R}^{d\times d}$.

Given hyperparameter $\kappa > 0$, the objective can be written

$$
\begin{aligned}
\hat{\mathcal{L}}_{\mathrm{BT}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) &:= \frac{1}{2} \sum_{i=1}^{d} \left((\hat{\boldsymbol{\Sigma}}_{AB})_{i,i} - 1 \right)^2 + \kappa \|\bar{\boldsymbol{\Sigma}}_{AB}\|_{\mathrm{F}}^2 + \iota_{\{\mathbf{I}\}}(\bar{\boldsymbol{\Sigma}}_{AA}) + \iota_{\{\mathbf{I}\}}(\bar{\boldsymbol{\Sigma}}_{BB}) \\
&= \underbrace{-\mathbf{Tr}(\hat{\boldsymbol{\Sigma}}_{AB}) + \kappa \|\bar{\boldsymbol{\Sigma}}_{AB}\|_{\mathrm{F}}^2}_{\text{covariance}} + \underbrace{\sum_{i=1}^{d}(\hat{\boldsymbol{\Sigma}}_{AB})_{i,i}^2 + \frac{d}{2} + \iota_{\{\mathbf{I}\}}(\bar{\boldsymbol{\Sigma}}_{AA}) + \iota_{\{\mathbf{I}\}}(\bar{\boldsymbol{\Sigma}}_{BB})}_{\text{variance regularization}} \, .
\end{aligned}
$$

Thus, this objective falls into the class as well, as the penalties enforce a particular variance structure akin to the regularizers above.

Example 3: Spectral Contrastive Loss. Finally, we consider the spectral contrastive loss from the pioneering work of HaoChen et al. (2021). This relates to similar viewpoints of contrastive learning as spectral methods found in the literature, such as the Laplacian eigenmap viewpoint of VICReg (Balestriero and LeCun, 2022, Section 3), the multidimensional scaling viewpoint of InfoNCE (Balestriero and LeCun, 2022, Section 4), or the recent spectral clustering viewpoint of SimCLR/CLIP (Tan et al., 2024, Sections 3 and 4). Recall that $\bar{\boldsymbol{\alpha}} := \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\alpha}(\boldsymbol{x}_i)$ and $\bar{\boldsymbol{\beta}} := \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\beta}(\boldsymbol{z}_i)$. In the multimodal setting, this loss (HaoChen et al., 2021, Eq. (6)) can be written as

$$
\begin{aligned}
\hat{\mathcal{L}}_{\mathrm{SC}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) &:= -\frac{1}{n} \sum_{i=1}^{n} \langle \boldsymbol{\alpha}(\boldsymbol{x}_i), \boldsymbol{\beta}(\boldsymbol{z}_i) \rangle + \frac{1}{n(n-1)} \sum_{i \neq j} \left(\langle \boldsymbol{\alpha}(\boldsymbol{x}_i), \boldsymbol{\beta}(\boldsymbol{z}_j) \rangle \right)^2 \\
&= -\frac{1}{n} \sum_{i=1}^{n} \langle \boldsymbol{\alpha}(\boldsymbol{x}_i) - \bar{\boldsymbol{\alpha}}, \boldsymbol{\beta}(\boldsymbol{z}_i) - \bar{\boldsymbol{\beta}} \rangle - \langle \bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}} \rangle \\
&\quad + \frac{1}{n(n-1)} \sum_{i \neq j} \left(\langle \boldsymbol{\alpha}(\boldsymbol{x}_i) - \bar{\boldsymbol{\alpha}}, \boldsymbol{\beta}(\boldsymbol{z}_j) - \bar{\boldsymbol{\beta}} \rangle \right)^2 + \left(\langle \bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}} \rangle \right)^2 \\
&= -\mathbf{Tr}(\hat{\boldsymbol{\Sigma}}_{AB}) + \frac{1}{n-1} \|\bar{\boldsymbol{\Sigma}}_{AB}\|_{\mathrm{F}}^2 - \langle \bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}} \rangle + \left(\langle \bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}} \rangle \right)^2 \\
&= \underbrace{-\mathbf{Tr}(\hat{\boldsymbol{\Sigma}}_{AB}) + \frac{1}{n-1} \|\bar{\boldsymbol{\Sigma}}_{AB}\|_{\mathrm{F}}^2}_{\text{covariance}} + \underbrace{\left(\langle \bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}} \rangle - \frac{1}{2} \right)^2 - \frac{1}{4}}_{\text{variance regularization}}
\end{aligned}
$$

where we set $\kappa := 1/(n-1)$ to complete the argument.

Example 4: Multimodal VICReg. We use a variant of the VICReg objective shown in Shwartz-Ziv et al. (2023, Equation 1). Note that this method is typically designed for one encoder being applied to two augmentations of the same object; however, it naturally generalizes to the multimodal case. The similarity graph simply connects paired observations, leading to the invariance term below. The multimodal VICReg objective has hyperparameters (c_1, c_2, c_3, κ). To state it, define the real-valued function $r(x) := \max\{0, c_1 - \sqrt{x + c_2}\}$ for $x \in \mathbb{R}$. We will also apply r to a matrix, which returns the matrix of element-wise applications of the function. The objective is written

$$
\hat{\mathcal{L}}_{\mathrm{VICReg}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{c_3}{2d} \underbrace{\left[\mathbf{Tr}\left(r(\hat{\boldsymbol{\Sigma}}_{AA}) \right) + \mathbf{Tr}\left(r(\hat{\boldsymbol{\Sigma}}_{BB}) \right) \right]}_{\text{variance}} + \frac{1}{2n} \underbrace{\|\mathbf{A} - \mathbf{B}\|_{\mathrm{F}}^2}_{\text{invariance}} + \underbrace{\kappa \|\bar{\boldsymbol{\Sigma}}_{AB}\|_{\mathrm{F}}^2}_{\text{covariance}} \, .
$$

While usually thought of as capturing a separate property, we will incorporate the invariance term into the other two terms, which crucially relies on having an extra degree of freedom via the second encoder (as opposed to the single-modality

setting). Define $\bar{\boldsymbol{\alpha}} := \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\alpha}(\boldsymbol{x}_i)$ and $\bar{\boldsymbol{\beta}} := \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\beta}(\boldsymbol{z}_i)$, then write

$$
\begin{aligned}
\frac{1}{2n}\|\mathbf{A}-\mathbf{B}\|_{\mathrm{F}}^2 &= \frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{\alpha}(\boldsymbol{x}_i)-\boldsymbol{\beta}(\boldsymbol{z}_i)\|_2^2 \\
&= \frac{1}{2n}\sum_{i=1}^{n}\|\boldsymbol{\alpha}(\boldsymbol{x}_i)-\bar{\boldsymbol{\alpha}}-\boldsymbol{\beta}(\boldsymbol{z}_i)+\bar{\boldsymbol{\beta}}\|_2^2 + \frac{1}{2}\|\bar{\boldsymbol{\alpha}}-\bar{\boldsymbol{\beta}}\|_2^2 \\
&= \frac{1}{2}\mathbf{Tr}\left(\hat{\boldsymbol{\Sigma}}_{AA}\right) + \frac{1}{2}\mathbf{Tr}\left(\hat{\boldsymbol{\Sigma}}_{BB}\right) - 2\,\mathbf{Tr}\left(\hat{\boldsymbol{\Sigma}}_{AB}\right) + \frac{1}{2}\|\bar{\boldsymbol{\alpha}}-\bar{\boldsymbol{\beta}}\|_2^2.
\end{aligned}
$$

The final term $\|\bar{\boldsymbol{\alpha}}-\bar{\boldsymbol{\beta}}\|_2^2$ can harmlessly be dropped in the objective, as all other terms do not depend on the individual means. Thus, we can redefine our VICReg objective as

$$
\hat{\mathcal{L}}_{\mathrm{VICReg}}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \underbrace{-\mathbf{Tr}\left(\hat{\boldsymbol{\Sigma}}_{AB}\right) + \kappa\|\bar{\boldsymbol{\Sigma}}_{AB}\|_{\mathrm{F}}^2}_{\text{covariance}}
$$

$$
+ \underbrace{\frac{1}{2}\left(\mathbf{Tr}\left(\hat{\boldsymbol{\Sigma}}_{AA}\right)+\mathbf{Tr}\left(\hat{\boldsymbol{\Sigma}}_{BB}\right)\right) + \frac{c_3}{2d}\left[\mathbf{Tr}\left(r(\hat{\boldsymbol{\Sigma}}_{AA})\right) + \mathbf{Tr}\left(r(\hat{\boldsymbol{\Sigma}}_{BB})\right)\right]}_{\text{variance regularization}},
$$

as intended.

F. Experimental Details

This appendix accompanies Sec. 4 with further details of the study. Before describing the experiments, we comment one quantity appearing in the risk bounds that is not analyzed experimentally is the distribution shift error that passes $\mathbf{L}^2(P_X)$-norm to the $\mathbf{L}^2(Q_X)$-norm from Sec. 3. For this, we refer the reader to the host of empirical work at the intersection of FSL, attribute-based and prompting-based ZSP, and distribution shift (see Recht et al. (2019); Hendrycks and Dietterich (2019); Goyal et al. (2023) and references therein).

F.1. Compute Environment

Experiments were run on a CPU/GPU workstation with 12 virtual cores, 126G of memory, and four NVIDIA TITAN Xp GPUs with 12G memory each. The code was written in Python 3.10 with the environment given by the YAML file in the supplement. The OpenCLIP and CLIP Benchmark repositories were either used directly or adapted in our codebase.

F.2. Evaluation Datasets

We use the following datasets as evaluation benchmarks for zero-shot image classification. Note that the following standard statistics describe their *test* sets.

- **Describable Textures Dataset (DTD):** 1,880 examples labeled with 47 classes (Cimpoi et al., 2014).

- **Flowers 102:** 6,149 examples labeled with 102 classes. (Nilsback and Zisserman, 2008).

- **FGVC Aircraft:** 3,333 examples labeled with 100 classes (Maji et al., 2013).

- **SUN397:** 21,750 examples labeled with 397 classes (Xiao et al., 2010).

- **ImageNet-1k:** 100,000 examples labeled with 998 classes. (Deng et al., 2009).

The **ImageNet-Captions** dataset (Fang et al., 2023) is also used for evaluation using a subset of 134,593 examples, whereas a 40,000 held-out subset is used to estimate the conditional means of the text embeddings. For the subsets of ImageNet-Captions, the exact filenames of the ImageNet-1k subsets are provided along with their captions. Image preprocessing for evaluation was done using the transformations in the PyTorch `transforms` module that were associated with each OpenCLIP model.

For the experiment behind Fig. 3, we design three in-distribution sub-tasks by randomly selecting collections of 50 classes $(\mathcal{Y}_1, \mathcal{Y}_2, \mathcal{Y}_3)$ from each of 998 classes, reserving held-out prompting examples $(Z_1, Y_1), \ldots, (Z_{15,000}, Y_{15,000})$, 100 for each of 150 classes. Then, for task i, using M examples $j_1(\boldsymbol{y}), \ldots, j_M(\boldsymbol{y})$ selected randomly without replacement for $\boldsymbol{y} \in \mathcal{Y}_i$, we use the vector $\frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\beta}(Z_{j_m(\boldsymbol{y})})$ as the class embedding (projected to unit norm). Using an evaluation set of approximately 25,000 examples from each sub-task, we compute the classification accuracy of this approach.

F.3. Model Specification and Hyperparameters

CLIP Architectures   First, we specify which OpenCLIP models and pre-training sets were used. These models were chosen due to their range of top-1 zero-shot accuracies on the ImageNet-1k benchmark (as shown below). As opposed to already highly performant models ($\geq$50% on ImageNet-1k), these models benefited more from optimized prompting techniques in our initial experiments.

Model	OpenCLIP Model Tag	Pre-Training Set Tag	ImageNet-1k Top-1 Acc.
ResNet-50	`RN50`	`yfcc15m`	28.11%
NLLB-CLIP	`nllb-clip-base`	`v1`	33.51%
ViT-B/32	`ViT-B-32`	`datacomp_m_s128m_b4k`	32.81%

Prompt-Generating Model   We employed the `meta-llama/Llama-3.2-1B-Instruct` model publicly available on HuggingFace. For the purpose of generation, we used a **top-**p hyperparameter of **0.9** and **temperature** hyperparameter of **0.99** for more diverse responses. Meta-prompting was based on the following instructions per dataset, which are slight variations of those used in Pratt et al. (2023):

- **Describable Textures Dataset (DTD):**
    - *"What does ___ material look like?"*,
    - *"What does a ___ surface look like?"*,
    - *"What does a ___ texture look like?"*,
    - *"What does a ___ object look like?"*,
    - *"What does a ___ pattern look like?"*

- **Flowers 102:**
    - *"Describe how to identify a(n) ___, a type of flower."*,
    - *"What does a(n) ___ flower looks like?"*

- **FGVC Aircraft:**
    - *"Describe a(n) ___ aircraft."*,
    - *"Describe the ___ aircraft."*

- **SUN397:**
    - *"Describe what a(n) ___ looks like."*,
    - *"How can you identify a(n) ___?"*,
    - *"Describe a photo of a(n) ___."*,
    - *"Describe the scene of a(n) ___."*

- **ImageNet-1k:**
    - *"Describe what a(n) ___ looks like."*,
    - *"How can you identify a(n) ___?"*,
    - *"What does a(n) ___ look like?"*,
    - *"Describe an image from the Internet of a(n) ___."*,
    - *"Write a caption of an image of a(n) ___."*

The following additional instruction was appended for better-formatted responses: *"Please format your response as one that contains only lower case letters and no special characters (including new lines, bold, and any markdown artifacts) other than a period ('.') or commas (','). The response should be a single sentence ending in a period that is directed toward the final instruction in this message. Your sentence should be a minimum of three words and a maximum of thirty."*.

Our reproducibility effort includes not only the full list of all 164,400 prompts generated from LlaMA 3, but the subset of prompts used for each class and each seed used to generate the figures in Sec. 4.

F.4. Derivation of Simulation Setting

The data-generating process for (X, Z, Y) in the simulation from Sec. 4 is as follows. Because we isolate the effect residual dependence in this simulation, we construct a joint distribution $P_{X,Y,Z}$ that satisfies Asm. 1, and moreover, such that $Q_{X,Z} = P_{X,Z}$ and $\rho_{Y,Z} = P_{Y,Z}$. Let $\mathcal{Y} = \{0, 1\}$, indicating binary classification. We consider $\mathcal{X} = \mathcal{Z} = \mathbb{R}^d$ and a pair of Gaussian distributions $(P_{X,Z|Y=0}, P_{X,Z|Y=1})$, where

$$\begin{bmatrix} X \\ Z \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_{X|\boldsymbol{y}} \\ \boldsymbol{\mu}_{Z|\boldsymbol{y}} \end{bmatrix}, \begin{bmatrix} \mathbf{C}_{XX|\boldsymbol{y}} & \mathbf{C}_{XZ|\boldsymbol{y}} \\ \mathbf{C}_{ZX|\boldsymbol{y}} & \mathbf{C}_{ZZ|\boldsymbol{y}} \end{bmatrix} \right) \text{ given } Y = \boldsymbol{y}. \tag{99}$$

Then, the distribution is fully specified by mean vectors and covariance matrices along with the parameter $p = \mathbb{P}[Y = 1]$. Letting $\mathcal{N}(\cdot; \boldsymbol{\mu}, \mathbf{C})$ indicate the density function of the $\mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ distribution, the *direct* predictor is equal to

$$p(\boldsymbol{x}) := \mathbb{E}_{P_{X,Y}}[Y|X](\boldsymbol{x}) = \frac{p\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{X|1})}{p\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{X|1}) + (1-p)\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{X|0})}. \tag{100}$$

Similarly, the *indirect predictor* is given by

$$\eta_\rho(\boldsymbol{x}) = \mathbb{E}_{P_{X,Z}}\left[\mathbb{E}_{P_{Z,Y}}[Y|Z]|X\right](\boldsymbol{x}) = \mathbb{E}_{P_{X,Z}}[p(Z)|X](\boldsymbol{x}), \tag{101}$$

$$p(\boldsymbol{z}) = \frac{p\mathcal{N}(Z; \boldsymbol{\mu}_{Z|1})}{p\mathcal{N}(Z; \boldsymbol{\mu}_{Z|1}) + (1-p)\mathcal{N}(Z; \boldsymbol{\mu}_{Z|0})}. \tag{102}$$

The expectation in (101) over Z given $X = x$ can be evaluated via simulation based on the mixture model $P_{Z|X=\boldsymbol{x}} = (1 - p(\boldsymbol{x}))P_{Z|X=\boldsymbol{x},Y=0} + p(\boldsymbol{x})P_{Z|X=\boldsymbol{x},Y=1}$ and the exact calculation

$$Z \sim \mathcal{N}\left(\boldsymbol{\mu}_{Z|\boldsymbol{y}} + \mathbf{C}_{ZX|\boldsymbol{y}}\mathbf{C}_{XX|\boldsymbol{y}}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{X|\boldsymbol{y}}), \mathbf{C}_{ZZ|\boldsymbol{y}} - \mathbf{C}_{ZX|\boldsymbol{y}}\mathbf{C}_{XX|\boldsymbol{y}}^{-1}\mathbf{C}_{XZ|\boldsymbol{y}} \right) \text{ given } X = \boldsymbol{x}, Y = \boldsymbol{y}.$$

Finally, the *residual dependence* $\mathbb{E}_{P_Z}[I(X;Y|Z)]$ can be computed by the following steps. First, notice that the conditional distribution of X given $Z = \boldsymbol{z}$ and $Y = \boldsymbol{y}$ is given by

$$X \sim \mathcal{N}\left(\boldsymbol{\mu}_{X|\boldsymbol{y}} + \mathbf{C}_{XZ|\boldsymbol{y}}\mathbf{C}_{ZZ|\boldsymbol{y}}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}_{Z|\boldsymbol{y}}), \mathbf{C}_{XX|\boldsymbol{y}} - \mathbf{C}_{XZ|\boldsymbol{y}}\mathbf{C}_{ZZ|\boldsymbol{y}}^{-1}\mathbf{C}_{ZX|\boldsymbol{y}} \right).$$

The likelihood ratio $\mathsf{S}_{\boldsymbol{z}}$ from (13) can be computed (where the evaluation at $\boldsymbol{x}$ refers to the density) via

$$\begin{aligned}
\mathsf{S}_{\boldsymbol{z}}(\boldsymbol{x}, \boldsymbol{y}) &= \frac{P_{X|Y=\boldsymbol{y},Z=\boldsymbol{z}}(\boldsymbol{x})[\boldsymbol{y}p(\boldsymbol{z}) + (1-\boldsymbol{y})(1-p(\boldsymbol{z}))]}{P_{X|Z=\boldsymbol{z}}(\boldsymbol{x})[\boldsymbol{y}p(\boldsymbol{z}) + (1-\boldsymbol{y})(1-p(\boldsymbol{z}))]} \\
&= \frac{P_{X|Y=\boldsymbol{y},Z=\boldsymbol{z}}(\boldsymbol{x})}{P_{X|Z=\boldsymbol{z}}(\boldsymbol{x})} \\
&= \frac{P_{X|Y=\boldsymbol{y},Z=\boldsymbol{z}}(\boldsymbol{x})}{(1 - p(\boldsymbol{z}))P_{X|Y=0,Z=\boldsymbol{z}}(\boldsymbol{x}) + p(\boldsymbol{z})P_{X|Y=1,Z=\boldsymbol{z}}(\boldsymbol{x})}.
\end{aligned}$$

To simulate from the marginal P_Z, we use the mixture $pP_{Z|Y=1} + (1-p)P_{Z|Y=0}$, after which (14) can be directly applied.

To interpolate between the setting in which $X \perp\!\!\!\perp Z|Y$ (the indirect predictor performs at chance) and $X \perp\!\!\!\perp Y|Z$ (the

indirect predictor is equivalent to the direct one), we use the setting

$$\boldsymbol{\mu}_{X|0} = \tfrac{1}{2}\mathbf{1}, \quad \boldsymbol{\mu}_{X|1} = -\tfrac{1}{2}\mathbf{1}.$$

Let $a, b > 0$ be constants and let $\theta \in [0, 1]$ be a parameter. Then, we define

$$\boldsymbol{\mu}_{Z|0} = 2\theta a \boldsymbol{\mu}_{X|0}, \quad \boldsymbol{\mu}_{Z|1} = 2\theta b \boldsymbol{\mu}_{X|1}$$

$$\mathbf{C}_{ZZ|0} = a\mathbf{I}, \quad \mathbf{C}_{ZX|0} = \tfrac{\theta a}{2}\mathbf{I}, \quad \mathbf{C}_{ZZ|1} = b\mathbf{I}, \quad \mathbf{C}_{ZX|1} = \tfrac{\theta b}{2}\mathbf{I}$$

and finally $\mathbf{C}_{XX|0} = (1 + \tfrac{a}{4})\mathbf{I}$ and $\mathbf{C}_{XX|1} = (1 + \tfrac{b}{4})\mathbf{I}$. Due to Gaussianity, it is clear that

$$\theta = 0 \implies \mathbf{C}_{ZX|0} = \mathbf{C}_{ZX|1} = \mathbf{0} \implies X \perp\!\!\!\perp Z | Y = y \,\forall y.$$

On the other hand, using the distribution of X given (Z, Y), that is,

$$X \sim \mathcal{N}\left(\boldsymbol{\mu}_{X|y} + \mathbf{C}_{XZ|y}\mathbf{C}_{ZZ|y}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}_{Z|y}), \mathbf{C}_{XX|y} - \mathbf{C}_{XZ|y}\mathbf{C}_{ZZ|y}^{-1}\mathbf{C}_{ZX|y}\right) \text{ given } Z = \boldsymbol{z}, Y = y,$$

we have that

$$\theta = 1 \implies \begin{cases} \boldsymbol{\mu}_{X|0} - \mathbf{C}_{XZ|0}\mathbf{C}_{ZZ|0}^{-1}\boldsymbol{\mu}_{Z|0} = \boldsymbol{\mu}_{X|1} - \mathbf{C}_{XZ|1}\mathbf{C}_{ZZ|1}^{-1}\boldsymbol{\mu}_{Z|1} \\ \mathbf{C}_{XZ|0}\mathbf{C}_{ZZ|0}^{-1} = \mathbf{C}_{XZ|1}\mathbf{C}_{ZZ|1}^{-1} \\ \mathbf{C}_{XX|0} - \mathbf{C}_{XZ|0}\mathbf{C}_{ZZ|0}^{-1}\mathbf{C}_{ZX|0} = \mathbf{C}_{XX|1} - \mathbf{C}_{XZ|1}\mathbf{C}_{ZZ|1}^{-1}\mathbf{C}_{ZX|1} \end{cases} \implies X \perp\!\!\!\perp Y | Z = \boldsymbol{z} \,\forall\boldsymbol{z}, \quad (103)$$

as the distribution of X remains the same given either $Z = \boldsymbol{z}, Y = 0$ or $Z = \boldsymbol{z}, Y = 1$. Thus, in the simulation, we interpolate between 0 and 1 for the value of θ. We set the parameters $a = 5$ and $b = 6$ simply to be different numbers for which $\mathbb{E}_{P_Z}[I(X; Y|Z)]$ can be computed in a numerically stable manner. We set $p = \tfrac{1}{2}$ and $d = 2$.

Finally, the lines labeled *CLIP* and *VICReg* in Fig. 2 indicate the predictors generated by training two MLP encoders using the corresponding objective on observations $\{(X_i, Z_i)\}_{i=1}^N$ for $N = 10,000$ pre-training observations. The encoder had a single hidden layer of 16 units and an output dimension of d. When performing classification, the prompting distribution used for the methods based on self-supervised learning is the true distribution of $Z|Y = \boldsymbol{y}$ with $M = 500$ samples, allowing us to isolate residual dependence while incurring no prompt bias and negligible prompt variance. Each model was trained for 30 epochs with the AdamW optimizer at a learning rate of 0.01. In the case of the VICReg objective, we used the parameterization of the original paper (Bardes et al., 2022) with the settings $(\gamma, \lambda, \mu, \nu, \epsilon) = (1, 25, 25, 1, 0.0001)$ as per the authors' recommendations (see their Eq. (6)).