Approximation in Contrastive Representation Learning

Yuanfan Li
School of mathematics and statistics
University of Sydney
yuli0846@uni.sydney.edu.au

Zihan Zhang
School of mathematics and statistics
University of Sydney
zihan.zhang@sydney.edu.au

Yiming Ying *
School of mathematics and statistics
University of Sydney
yiming.ying@sydney.edu.au

Ding-Xuan Zhou
School of mathematics and statistics
University of Sydney
dingxuan.zhou@sydney.edu.au

Abstract

Contrastive representation learning (CRL) is a fundamental paradigm for learning transferable representations. Recent work identifies its population target as the pointwise mutual information (PMI) $\log \frac{p(\mathbf{x},\mathbf{y})}{p_{\mathcal{X}}(\mathbf{x})p_{\mathcal{Y}}(\mathbf{y})}$ between two modalities $\mathcal{X}, \mathcal{Y}$, where p is the joint density function and $p_{\mathcal{X}}, p_{\mathcal{Y}}$ are the marginals. However, it remains unclear why standard inner-product scores $f(\mathbf{x})^\top g(\mathbf{y})$ can effectively approximate a generally nonseparable population target PMI. In this paper, we address this question from the aspect of approximation for CRL. By linking PMI to a compact operator, we show that D-dimensional inner-product models perform rank-D spectral approximation, with error controlled by the spectral tail. We then study neural network realizations: in general settings, we show that deep neural network scoring functions are universally consistent, and the approximation error rates depend on the smoothness of the PMI function; under structured low-rank data models we obtain improved rates attained by moderate-width encoders. Finally, we establish a fast calibration rate from contrastive excess risk to downstream retrieval error under a density-ratio Tsybakov noise condition.

1 Introduction

Representation learning has been a prominent paradigm in foundation models and achieved great success [Radford et al., 2021, Jia et al., 2021, Karpukhin et al., 2020]. Typically, representations are learned from large-scale datasets and then transferred to downstream tasks. One of the most successful approaches is *contrastive representation learning* (CRL), which has been widely used in computer vision, natural language processing, and multimodal learning [Chen et al., 2020, He et al., 2020, An et al., 2023, Radford et al., 2021, Jia et al., 2021].

For two modalities $\mathcal{X}$ and $\mathcal{Y}$, CRL aims to learn a scoring function $s : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$, that assigns higher scores to positive pairs than to negative pairs. By pulling together semantically related pairs $(\mathbf{x}, \mathbf{y})$ and pushing apart mismatched pairs $(\mathbf{x}, \mathbf{y}')$ in the embedding space, CRL learns transferable and task-agnostic representations. In the self-supervised setting, this mechanism is commonly implemented by minimizing the population objective [Wang and Isola, 2020]:

$$\mathcal{L}(s) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p}\, \tau \log \mathbb{E}_{\mathbf{y}'\sim p_{\mathcal{Y}}} \exp\left(\frac{s(\mathbf{x},\mathbf{y}') - s(\mathbf{x},\mathbf{y})}{\tau}\right), \tag{1}$$

*Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

where p is the joint distribution of $\mathbf{x}, \mathbf{y}$, $p_{\mathcal{Y}}$ is the marginal distribution of $\mathbf{y}$.

[Bao et al., 2022, Ash et al., 2022] analyse surrogate gaps and excess risks for contrastive losses, providing insights into their generalisation behaviour. However, they do not establish Fisher consistency or explicitly characterise the population minimiser of the contrastive objective.

To address this, Li et al. [2026] show that the population minimiser s^* of the contrastive loss indeed recovers the log-density ratio $\log \frac{p(\mathbf{x},\mathbf{y})}{p_{\mathcal{X}}(\mathbf{x})p_{\mathcal{Y}}(\mathbf{y})}$, which is the pointwise mutual information (PMI) of two modalities (up to an additive function). Building on this characterisation, they further derive a calibration-type inequality: for any scoring function s, there holds

$$\mathcal{E}^* - \mathcal{E}(s) \lesssim \sqrt{\mathcal{L}(s) - \mathcal{L}^*}, \tag{2}$$

which quantitatively connects pretraining excess risk to downstream retrieval performance. Here, the retrieval risk of a scoring function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is defined as

$$\mathcal{E}(s) := \mathbb{P}\big(s(\mathbf{x}, \mathbf{y}) \geq s(\mathbf{x}, \mathbf{y}')\big),$$

which measures the probability that a randomly drawn positive pair is ranked above a negative pair. The optimal (Bayes) retrieval risk is $\mathcal{E}^* := \max_s \mathcal{E}(s)$, where the maximum is taken over all measurable functions $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, and $\mathcal{L}^* = \mathcal{L}(s^*)$ is the minimal contrastive loss.

Typically in the context of CRL [Chen et al., 2020, Radford et al., 2021], the scoring function takes the form of inner product of representations , i.e., $s \in \mathcal{S}_{\mathrm{IP}} = \{s(\mathbf{x}, \mathbf{y}) = f(\mathbf{x})^\top g(\mathbf{y}) \mid f \in \mathcal{F}, g \in \mathcal{G}\}$, where $\mathcal{F}, \mathcal{G}$ are hypothesis spaces, e.g., deep neural networks (DNNs). This limits the class of admissible scoring functions, as there is no obvious way to decompose $\log \frac{p(\mathbf{x},\mathbf{y})}{p_{\mathcal{X}}(\mathbf{x})p_{\mathcal{Y}}(\mathbf{y})}$ into separate factors of $\mathbf{x}$ and $\mathbf{y}$, and thus direct inner-product approximation appears challenging. Yet, the learned representations still generalizes in downstream tasks. Building on this perspective, we ask the following central question:

> How well can a given inner-product representation class approximate pointwise mutual information underlying contrastive learning?

To address this question, we study CRL within the framework of statistical learning theory. Let $\hat{s}$ be a minimizer of an empirical loss $\widehat{\mathcal{L}}$ over a hypothesis class $\mathcal{S}_{\mathrm{IP}}$, and let $\bar{s} \in \arg\min_{s \in \mathcal{S}_{\mathrm{IP}}} \mathcal{L}(s)$. The excess risk admits the decomposition

$$\begin{aligned}\mathcal{L}(\hat{s}) - \mathcal{L}^* &= \big[\mathcal{L}(\hat{s}) - \widehat{\mathcal{L}}(\hat{s})\big] + \big[\widehat{\mathcal{L}}(\hat{s}) - \widehat{\mathcal{L}}(\bar{s})\big] + \big[\widehat{\mathcal{L}}(\bar{s}) - \mathcal{L}(\bar{s})\big] + \big[\mathcal{L}(\bar{s}) - \mathcal{L}^*\big] \\ &\leq 2 \sup_{s \in \mathcal{S}_{\mathrm{IP}}} |\mathcal{L}(s) - \widehat{\mathcal{L}}(s)| + \inf_{s \in \mathcal{S}_{\mathrm{IP}}} \big(\mathcal{L}(s) - \mathcal{L}^*\big).\end{aligned}$$

Existing analyses [Saunshi et al., 2019, Lei et al., 2023, Li et al., 2026] establish bounds on the generalization gap, i.e., $\sup_{s \in \mathcal{S}_{\mathrm{IP}}} |\mathcal{L}(s) - \widehat{\mathcal{L}}(s)|$, but do not address whether the inner-product hypothesis class $\mathcal{S}_{\mathrm{IP}}$ is expressive enough to capture the optimal log-density-ratio structure. In particular, the approximation error $\inf_{s \in \mathcal{S}_{\mathrm{IP}}} \big(\mathcal{L}(s) - \mathcal{L}^*\big)$ remains largely unexplored.

As the sample size increases, the generalization term vanishes, and the performance is ultimately governed by the *approximation error* $\inf_{s \in \mathcal{S}_{\mathrm{IP}}} \big(\mathcal{L}(s) - \mathcal{L}^*\big)$, which measures the gap between the best achievable risk within $\mathcal{S}_{\mathrm{IP}}$ and the population optimum. Characterizing this quantity is therefore key to explaining the empirical success of overparameterized models and understanding the fundamental limits of contrastive representation learning.

In this work, we develop a theoretical analysis of CRL that characterizes the role of approximation in determining its achievable performance. We also propose a density-ratio Tsybakov low-noise condition and establish a fast calibration rate. Our contributions are listed as follows:

- We first study the approximation capability of the finite-dimensional inner-product model class

$$\mathcal{S}_{\mathrm{IP}} = \mathcal{S}_D := \{\, s(\mathbf{x}, \mathbf{y}) = f(\mathbf{x})^\top g(\mathbf{y}) \mid f : \mathcal{X} \to \mathbb{R}^D,\, g : \mathcal{Y} \to \mathbb{R}^D \,\}.$$

  Let the *PMI operator* $T : L^2(P_{\mathcal{X}}) \to L^2(P_{\mathcal{Y}})$ be defined by $(Tf)(\mathbf{y}) := \int \log \frac{p(\mathbf{x},\mathbf{y})}{p_{\mathcal{X}}(\mathbf{x})p_{\mathcal{Y}}(\mathbf{y})}\, f(\mathbf{x})\, dP_{\mathcal{X}}(\mathbf{x})$. Denote its singular value decomposition by $T\psi_k =$

$\sigma_k\phi_k, \quad T^*\phi_k = \sigma_k\psi_k$, where $\{\psi_k\}$ and $\{\phi_k\}$ are orthonormal systems in $L^2(P_{\mathcal{X}})$ and $L^2(P_{\mathcal{Y}})$, respectively. We show that CRL implicitly performs a low-rank factorisation of T, where the encoders f and g capture the top-D singular functions $\{\psi_k, \phi_k\}_{k=1}^D$. The approximation error is governed by the representation dimension D and the decay of the singular values $\{\sigma_k\}$.

- We further study the approximation error when f, g are parameterized by DNNs. We first show the universal approximation property of DNNs. Then we derive an explicit rate under some smoothness assumptions of the PMI function. We further analyze a structured data model where the PMI operator is low-rank, for which we show that small dimensional encoders can achieve small approximation error through learning the singular functions $\{\psi_k\}, \{\phi_k\}$.
- We present fine-grained analysis of calibration-type inequality for retrieval in the case of low noise. We introduce a density ratio Tsybakov noise condition, formulated through the density-ratio margin between two samples. Large margin implies low-noise and easy ranking task, thus leading to tighter calibration inequality.

2 Related Work

The seminal work of Saunshi et al. [2019] proposed a theoretical framework for analyzing the generalization performance of contrastive representation learning, which was later refined by Lei et al. [2023]. Another line of work studies the *surrogate gap*, namely the discrepancy between the upstream contrastive objective and downstream task performance [Bao et al., 2022, Ash et al., 2022]. More recently, Li et al. [2026] further developed the statistical learning theory of CRL by establishing Fisher consistency, calibration inequalities, and improved generalization bounds.

A complementary line of research analyzes CRL from spectral and kernel perspectives. HaoChen et al. [2021] introduced a spectral contrastive loss and showed that it recovers eigenfunctions of a Laplacian operator defined on an augmentation graph. Saunshi et al. [2022] emphasized the importance of inductive bias in explaining the success of contrastive learning. Johnson et al. [2023] proposed the positive-pair kernel and connected CRL with principal component analysis, while Zhai et al. [2024] studied augmentation-based CRL through RKHS approximation and regression, showing that sufficiently strong augmentations alone can induce generalization. Our work differs from these analyses in two key aspects. First, prior works typically study matrices/operators induced by positive-pair distributions or augmentation graphs, which do not directly correspond to the population minimizer of the standard log-sum-exp contrastive objective. In contrast, we analyze the operator induced by the pointwise mutual information (PMI), which characterizes the optimal contrastive score. Second, we explicitly study the two-encoder structure $f(\mathbf{x})$ and $g(\mathbf{y})$ and its approximation properties under inner-product scoring functions.

Several recent studies have also begun to investigate approximation aspects of CRL. Oko et al. [2025] proposed a joint generative hierarchical model and considered a scoring function of the form $\tau^w(\mathrm{softmax}(f(\mathbf{x})), \mathrm{softmax}(g(\mathbf{y})))$, where τ^w is a nonlinear function parameterized by w. They showed that transformers can approximate the optimal score by implementing belief propagation. Uesaka et al. [2025] proposed weighted point-set embeddings with a universal kernel to compute pointwise mutual information. However, these approaches rely on more expressive scoring functions or kernel constructions, and do not fully explain why the standard inner-product form $s(\mathbf{x}, \mathbf{y}) = f(\mathbf{x})^\top g(\mathbf{y})$ is sufficiently expressive in practice.

Further related work is discussed in Appendix A.

3 Problem Formulation

Notations. Let $\mathbb{N}$ denote the set of positive integers, $d_1, d_2 \in \mathbb{N}$, and $\mathcal{X} \subset \mathbb{R}^{d_1}$ and $\mathcal{Y} \subset \mathbb{R}^{d_2}$. Denote $(\mathbf{x}, \mathbf{y})$ a data pair with $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$. We denote by $P_{\mathcal{X}\mathcal{Y}}$ the joint distribution of $(\mathbf{x}, \mathbf{y})$. The marginal distributions of $P_{\mathcal{X}\mathcal{Y}}$ on $\mathcal{X}$ and $\mathcal{Y}$ are denoted by $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$, respectively. If $P_{\mathcal{X}\mathcal{Y}}$ is continuous, we use $p(\mathbf{x}, \mathbf{y})$, $p_{\mathcal{X}}(\mathbf{x})$ and $p_{\mathcal{Y}}(\mathbf{y})$ to denote the probability density functions of $P_{\mathcal{X}\mathcal{Y}}$, $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$, respectively; if $P_{\mathcal{X}\mathcal{Y}}$ is discrete, we use $p(\mathbf{x}, \mathbf{y})$, $p_{\mathcal{X}}(\mathbf{x})$ and $p_{\mathcal{Y}}(\mathbf{y})$ to denote the probability mass functions of $P_{\mathcal{X}\mathcal{Y}}$, $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$, respectively. By default, we assume that $P_{\mathcal{X}\mathcal{Y}}$ is continuous. For simplicity, we sometimes write $\mathbb{E}_{\mathbf{x}}$ for $\mathbb{E}_{\mathbf{x}\sim p_{\mathcal{X}}}$. Let $L_2(P_{\mathcal{X}})$ and $L_2(P_{\mathcal{Y}})$ denote the Hilbert spaces of square-integrable functions with respect to $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$, respectively. We write

$a \lesssim b$ if $a \leq c\, b$ for some constant $c > 0$, and $a \asymp b$ if there exist constants $c_1, c_2 > 0$ such that $c_1 a \leq b \leq c_2 a$. Denote by $\mathbb{I}$ the indicator function, i.e., $\mathbb{I}\{E\} = 1$ if the event E holds true.

We focus on self-supervised CRL throughout the paper. Our theoretical analysis can be naturally extended to general distributions for positive and negative samples. Recall that the population contrastive risk is defined as follows [Wang and Isola, 2020, Li et al., 2026],

$$\mathcal{L}(s) \coloneqq \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p}\left[\tau \log \mathbb{E}_{\mathbf{y}'\sim p_{\mathcal{Y}}} \exp\left(\frac{s(\mathbf{x},\mathbf{y}') - s(\mathbf{x},\mathbf{y})}{\tau}\right)\right], \tag{3}$$

where $\tau > 0$ is the temperature parameter. This objective encourages the score of a positive pair $(\mathbf{x}, \mathbf{y})$ to be larger than that of a negative pair $(\mathbf{x}, \mathbf{y}')$.

Population Target as PMI. The optimal score for the contrastive risk (3) is established in the following lemma.

Lemma 3.1 ([Li et al., 2026])**.** *Let $\mathcal{L}(s)$ be defined as in* (3)*. A measurable function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ minimizes $\mathcal{L}(s)$ if and only if it satisfies $s(\mathbf{x},\mathbf{y}) = \tau \log \frac{p(\mathbf{y}|\mathbf{x})}{p_{\mathcal{Y}}(\mathbf{y})} + a(\mathbf{x})$ for some measurable function $a : \mathcal{X} \to \mathbb{R}$.*

Let $a(\mathbf{x}) = 0$, then the essential population target of CRL is the pointwise mutual information (PMI)

$$\log \frac{p(\mathbf{y} \mid \mathbf{x})}{p_{\mathcal{Y}}(\mathbf{y})} = \log \frac{p(\mathbf{x},\mathbf{y})}{p_{\mathcal{X}}(\mathbf{x}) p_{\mathcal{Y}}(\mathbf{y})} \coloneqq \mathrm{PMI}(\mathbf{x},\mathbf{y}). \tag{4}$$

Representation Classes and Approximation Error. In practice, PMI is learned through a structured hypothesis class. The most common choice is the inner-product form

$$\mathcal{S}_{\mathrm{IP}} = \left\{ s(\mathbf{x},\mathbf{y}) = f(\mathbf{x})^\top g(\mathbf{y}) \,\middle|\, f \in \mathcal{F},\ g \in \mathcal{G} \right\}. \tag{5}$$

where $\mathcal{F}$ and $\mathcal{G}$ are representation classes, such as deep neural network classes. The central object of this paper is to approximate PMI through the functions in $\mathcal{S}_{\mathrm{IP}}$. To this end, we analyze the approximation error

$$\inf_{s\in\mathcal{S}_{\mathrm{IP}}} \left[\mathcal{L}(s) - \mathcal{L}^*\right].$$

In this paper, we consider two representation classes: 1) a finite-dimensional *inner-product model class*, where the dimension D controls the representation capacity:

$$\mathcal{S}_{\mathrm{IP}} = \mathcal{S}_D \coloneqq \left\{ s(\mathbf{x},\mathbf{y}) = f(\mathbf{x})^\top g(\mathbf{y}) \,\middle|\, f : \mathcal{X} \to \mathbb{R}^D,\ g : \mathcal{Y} \to \mathbb{R}^D \text{ measurable} \right\}. \tag{6}$$

2) a hypothesis space $\mathcal{S}_{\mathrm{NN}}$ where $\mathcal{F}, \mathcal{G}$ are DNNs, which is defined in (12).

Calibration Inequality for Downstream Retrieval. We use the scoring function learned by CRL for downstream retrieval. For a scoring function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, we measure its retrieval quality by the AUC-type population criterion

$$\mathcal{E}(s) \coloneqq \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p}\mathbb{E}_{\mathbf{y}'\sim p_{\mathcal{Y}}}\left[\mathbb{I}\{s(\mathbf{x},\mathbf{y}) > s(\mathbf{x},\mathbf{y}')\} + \frac{1}{2}\mathbb{I}\{s(\mathbf{x},\mathbf{y}) = s(\mathbf{x},\mathbf{y}')\}\right]. \tag{7}$$

Larger values of $\mathcal{E}(s)$ correspond to better retrieval performances. We define the optimal retrieval value as $\mathcal{E}^* \coloneqq \max_s \mathcal{E}(s)$, where the maximum is taken over all measurable functions $s : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$. It is expected that minimizing the contrastive risk $\mathcal{L}(s)$ would maximize $\mathcal{E}(s)$. This statistical property, can be quantitatively described in the following inequality [Li et al., 2026]:

$$\mathcal{E}^* - \mathcal{E}(s) \lesssim \sqrt{\mathcal{L}(s) - \mathcal{L}^*}. \tag{8}$$

This shows that a small contrastive excess risk is guaranteed to achieve a strong retrieval performance in the downstream task. Therefore, the approximation error controls the best downstream retrieval performance achievable by $\mathcal{S}_{\mathrm{IP}}$: $\inf_{s\in\mathcal{S}_{\mathrm{IP}}}\left[\mathcal{E}^* - \mathcal{E}(s)\right] \lesssim \sqrt{\inf_{s\in\mathcal{S}_{\mathrm{IP}}}\left[\mathcal{L}(s) - \mathcal{L}^*\right]}$.

4 Approximation Analysis with Finite-dimensional Inner-Product Representations

In this section we investigate how well the pointwise mutual information (PMI) can be approximated by D-dimensional inner-product scores $\mathcal{S}_D$ defined in (6). The analysis constructs low-rank spectral comparators for a PMI operator, with approximation guarantees controlled by two factors: the embedding dimension D and the spectral decay of the PMI operator.

Matrix factorization for discrete spaces. When $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ are finite spaces, the PMI function can be viewed as a matrix $P \in \mathbb{R}^{n \times m}$ with entries

$$P_{ij} \coloneqq \log \frac{p(\mathbf{x}_i, \mathbf{y}_j)}{p_{\mathcal{X}}(\mathbf{x}_i) p_{\mathcal{Y}}(\mathbf{y}_j)}.$$

For two encoders $f(\mathbf{x}_i), g(\mathbf{y}_j) \in \mathbb{R}^D$, let $F = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)]^\top \in \mathbb{R}^{n \times D}$ and $G = [g(\mathbf{y}_1), \ldots, g(\mathbf{y}_m)]^\top \in \mathbb{R}^{m \times D}$, then a rank-at-most-$D$ factorization $P \approx FG^\top$ provides a comparator for inner-product contrastive scores (with score scale τ). By the Eckart–Young–Mirsky theorem [Eckart and Young, 1936], the best such approximation in the Frobenius norm is given by the truncated SVD: if $P = U\Sigma V^\top$, then $P_D = U_D \Sigma_D V_D^\top$, where U_D is the first D columns of U, V_D is the first D columns of V, Σ_D is the upper left block of Σ, and one may take $F = U_D \Sigma_D^{1/2}$, $G = V_D \Sigma_D^{1/2}$.

Spectral decomposition on continuous domains. The matrix viewpoint extends naturally to continuous spaces. Assume the PMI is square-integrable, $\mathrm{PMI}(\mathbf{x}, \mathbf{y}) \in L_2(P_{\mathcal{X}} \otimes P_{\mathcal{Y}})$. Define the *PMI operator* $T : L_2(P_{\mathcal{X}}) \to L_2(P_{\mathcal{Y}})$ by

$$(Tf)(\mathbf{y}) = \int_{\mathcal{X}} \mathrm{PMI}(\mathbf{x}, \mathbf{y})\, f(\mathbf{x})\, dP_{\mathcal{X}}(\mathbf{x}). \tag{9}$$

Its adjoint is $T^* g(\mathbf{x}) = \int_{\mathcal{Y}} \mathrm{PMI}(\mathbf{x}, \mathbf{y}) g(\mathbf{y}) dP_{\mathcal{Y}}(\mathbf{y})$. The compositions T^*T and TT^* are integral operators with kernels

$$K_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') = \mathbb{E}_{\mathbf{y} \sim p_{\mathcal{Y}}}\big[\mathrm{PMI}(\mathbf{x}, \mathbf{y})\, \mathrm{PMI}(\mathbf{x}', \mathbf{y})\big], \qquad K_{\mathcal{Y}}(\mathbf{y}, \mathbf{y}') = \mathbb{E}_{\mathbf{x} \sim p_{\mathcal{X}}}\big[\mathrm{PMI}(\mathbf{x}, \mathbf{y})\, \mathrm{PMI}(\mathbf{x}, \mathbf{y}')\big].$$

$K_{\mathcal{X}}$ measures similarity between two anchors through the lens of the entire $\mathcal{Y}$ modality; for instance, two images are close if they share similar textual descriptions.

Since the kernel is square-integrable, T is a Hilbert–Schmidt operator. The spectral theorem then yields the following decomposition (proof in Appendix B.1).

Lemma 4.1 (Spectral decomposition of the PMI)**.** *T^*T and TT^* share the same non-zero eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \searrow 0$. There exist orthonormal systems $\{\psi_i\} \subset L_2(P_{\mathcal{X}})$ and $\{\phi_i\} \subset L_2(P_{\mathcal{Y}})$ such that $T\psi_i = \sigma_i \phi_i,\ T^*\phi_i = \sigma_i \psi_i$ with $\sigma_i = \sqrt{\lambda_i}$. Moreover,*

$$\mathrm{PMI}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{\infty} \sigma_i\, \psi_i(\mathbf{x})\, \phi_i(\mathbf{y}) \quad \text{in } L_2(P_{\mathcal{X}} \otimes P_{\mathcal{Y}}). \tag{10}$$

Furthermore, the ranges $\mathrm{Im}(T)$ and $\mathrm{Im}(T^*)$ are exactly the reproducing kernel Hilbert spaces (RKHSs) induced by $K_{\mathcal{Y}}$ and $K_{\mathcal{X}}$, respectively. We denote these spaces by $\mathcal{H}_{\mathcal{Y}} := \mathrm{Im}(T)$ and $\mathcal{H}_{\mathcal{X}} := \mathrm{Im}(T^*)$, equipped with their RKHS norms.

(10) is the infinite-dimensional analogue of the SVD of the PMI matrix. Spectral truncation therefore provides explicit encoder coordinates through the singular functions $\{\psi_i\}$ and $\{\phi_i\}$.

Approximation error analysis. If the PMI operator has finite rank $K < \infty$ (i.e., $\lambda_i = 0$ for $i > K$), then encoders of dimension K suffice to represent the optimal score exactly: $s^*(\mathbf{x}, \mathbf{y}) = \tau \sum_{i=1}^{K} \sigma_i\, \psi_i(\mathbf{x}) \phi_i(\mathbf{y})$ can be written as $f(\mathbf{x})^\top g(\mathbf{y})$ with $f(\mathbf{x}) = (\sqrt{\tau}\sigma_1^{1/2}\psi_1(\mathbf{x}), \ldots, \sqrt{\tau}\sigma_K^{1/2}\psi_K(\mathbf{x}))$ and $g(\mathbf{y}) = (\sqrt{\tau}\sigma_1^{1/2}\phi_1(\mathbf{y}), \ldots, \sqrt{\tau}\sigma_K^{1/2}\phi_K(\mathbf{y}))$. Section 5.2 provides a low-rank data model where this finite-rank structure emerges.

In the infinite-rank case, when the encoders are restricted to a fixed dimension D, the optimal L_2-approximation of the PMI within $\mathcal{S}_D$ is indeed achieved by the truncated spectral decomposition $\sum_{i=1}^{D} \sqrt{\lambda_i}\, \psi_i(\mathbf{x})\, \phi_i(\mathbf{y})$.

To derive the approximation error, we introduce a regularity condition on the density ratio.

Assumption 4.2. There exists a constant $\alpha > 1$ such that, for all $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$, $\frac{p(\mathbf{x}, \mathbf{y})}{p_{\mathcal{X}}(\mathbf{x})\, p_{\mathcal{Y}}(\mathbf{y})} \leq \alpha$.

Note that we do not impose positive lower bound, which is employed in Oko et al. [2025]. Large α usually implies imbalanced distribution. Building on this assumption, we obtain the following bound (proofs are provided in Appendix B.2).

Theorem 4.3 (Approximation error of inner-product scores)**.** *Suppose that $\mathcal{X}, \mathcal{Y}$ be two compact domains, the PMI function $\log \frac{p(\mathbf{x},\mathbf{y})}{p_{\mathcal{X}}(\mathbf{x})p_{\mathcal{Y}}(\mathbf{y})}$ is continuous on $\mathcal{X} \times \mathcal{Y}$ and that Assumption 4.2 holds. Let $\mathcal{S}_D$ be as in* (6) *and assume the spectral score $s_D(\mathbf{x}, \mathbf{y}) = \tau \sum_{i=1}^{D} \sqrt{\lambda_i}\psi_i(\mathbf{x})\phi_i(\mathbf{y})$ satisfy $\|s_D\|_\infty \leq B$. Then*

$$\inf_{s \in \mathcal{S}_D} [\mathcal{L}(s) - \mathcal{L}(s^*)] \lesssim \sum_{i=D+1}^{\infty} \lambda_i.$$

Hence the tail sum of the eigenvalues $\{\lambda_i\}$ directly measures the intrinsic approximation hardness: a fast decay implies that a small encoder dimension already captures most of the mutual information. To illustrate this, we consider a case when the eigenvalues admit a polynomial decay [Ying and Pontil, 2008, Guo et al., 2023].

Corollary 4.4. *Under the conditions of Theorem 4.3, suppose the eigenvalues decay polynomially: $\lambda_i \asymp i^{-\gamma}$ for some $\gamma > 1$. Then*

$$\inf_{s \in \mathcal{S}_D} [\mathcal{L}(s) - \mathcal{L}(s^*)] \lesssim \frac{(D+1)^{1-\gamma}}{\gamma - 1}.$$

To achieve an approximation error ε, it suffices to choose $D \approx \varepsilon^{-1/(\gamma-1)}$. Larger γ (faster decay) thus permits much smaller embedding dimension. Smooth PMI kernels typically exhibit rapid eigenvalue decay [Ferreira and Menegatto, 2009]. We conduct numerical experiments on Gaussian mixture model in Appendix I.1.

Based on our theoretical findings, we give two insights on why moderate dimensions of representations work well in practice: (i) the data exhibits a low-dimensional semantic structure, (ii) the PMI function is smooth, meaning that similar inputs have similar co-occurrence patterns and outliers are rare.

Remark 4.5 (Contrastive risk versus L_2 approximation). In our analysis, we first give the approximation error in L_2 norm then we control the contrastive risk by L_2 distance. However, these two measures are not equivalent in general. The log-sum-exp structure of contrastive risk is more sensitive to hard negatives. Furthermore, contrastive excess risk can be written as KL divergence between the conditional distribution $p(\cdot|\mathbf{x})$ and a distribution induced by $s : q_s(\mathbf{y} \mid \mathbf{x}) = p_{\mathcal{Y}}(\mathbf{y})e^{s(\mathbf{x},\mathbf{y})/\tau}/\mathbb{E}_{\mathbf{y}'}e^{s(\mathbf{x},\mathbf{y}')/\tau}, \mathcal{L}(s) - \mathcal{L}^* = \tau\mathbb{E}_{\mathbf{x}}\mathrm{KL}(p(\cdot \mid \mathbf{x})\|q_s(\cdot \mid \mathbf{x}))$. Consequently, the global minimizer over the class $\mathcal{S}_D$ of D-dimensional inner-product scores is an KL-divergence projection.

Remark 4.6 (RKHS encoders and cosine similarity). We also study the approximation ability of RKHSs $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ with bounded norms. The first r spectral components have squared encoder norms $\tau \sum_{j=1}^{r} \lambda_j^{-1/2}$; RKHS with larger norm therefore permits a smaller spectral tail. For cosine scoring functions, we show that $D + 2$-dimension cosine representations can realize truncated scores with top-D eigenfunctions. See Appendix B.3 for more details.

5 Approximation Analysis with DNNs

The previous section characterises the approximation error of PMI via a spectral decomposition, indicating that inner-product spaces can approximate PMI well provided the eigenvalues of the operator decay fast. We now turn to the question of *realisability*: can PMI be effectively represented by practical models?

In practice, CRL parameterises the scoring function via inner products of deep neural network (DNN) encoders. In Section 5.1, we provide a general approximation analysis for DNN-based inner-product models, providing theoretical guarantees for their expressive power in approximating PMI. In Section 5.2, we consider a structured data model where the PMI operator is low-rank, we show that DNNs can overcome the curse of dimensionality by learning the singular functions of PMI.

5.1 Approximation via DNNs: General Case

We show that such DNN-based representations can well approximate the target PMI function $\log \frac{p(\mathbf{x},\mathbf{y})}{p_{\mathcal{X}}(\mathbf{x})p_{\mathcal{Y}}(\mathbf{y})}$, and we establish corresponding approximation error.

Neural network space. Let $\sigma(t) = \max\{0, t\}$ be the ReLU function. For any $d \in \mathbb{N}$ and any $\mathbf{v} \in \mathbb{R}^d$, we define the so-called $\mathbf{v}$-shifted ReLU function

$$\sigma_{\mathbf{v}} : \mathbb{R}^d \to [0,\infty)^d,\ \mathbf{x} \mapsto \big(\sigma((\mathbf{x}-\mathbf{v})_1), \sigma((\mathbf{x}-\mathbf{v})_2), \ldots, \sigma((\mathbf{x}-\mathbf{v})_d)\big).$$

A ReLU neural network is a function f of the form

$$f : \mathbb{R}^d \to \mathbb{R}^D,\ \mathbf{x} \mapsto \boldsymbol{W}_L \sigma_{\mathbf{v}_L} \boldsymbol{W}_{L-1} \sigma_{\mathbf{v}_{L-1}} \cdots \boldsymbol{W}_2 \sigma_{\mathbf{v}_2} \boldsymbol{W}_1 \sigma_{\mathbf{v}_1} \boldsymbol{W}_0 \mathbf{x}, \tag{11}$$

where $d \in \mathbb{N}$, $D \in \mathbb{N}$, $L \in \mathbb{N} \cup \{0\}$, $\mathrm{m}_i \in \mathbb{N}$ $(0 \le i \le L+1)$, $\mathrm{m}_0 = d$, $\mathrm{m}_{L+1} = D$, $\mathbf{v}_i \in \mathbb{R}^{\mathrm{m}_i}$ $(1 \le i \le L)$, $\boldsymbol{W}_i \in \mathbb{R}^{\mathrm{m}_{i+1} \times \mathrm{m}_i}$ $(0 \le i \le L)$. Define the space

$$\mathcal{F}_{\texttt{NN}}(d, L, N, D) := \{f \mid f \text{ is defined by (11) with } m_i \le N \text{ for } 1 \le i \le L\},$$

which consists of all ReLU DNNs of which the depth is L, the number of input units is d, the number of output units is D, and the number of units in each hidden layer is less than or equal to N.

We consider scoring functions $s(\mathbf{x}, \mathbf{y})$ which take the form of $s(\mathbf{x}, \mathbf{y}) = f(\mathbf{x})^\top g(\mathbf{y})$, where $f(\mathbf{x})$ and $g(\mathbf{y})$ are ReLU neural networks. The hypothesis space of such scoring functions is specified as

$$\begin{aligned}\mathcal{S}_{\mathrm{NN}} &= \mathcal{F}_{\texttt{S}}(L_1, N_1, L_2, N_2, D) \\ &:= \left\{ s(\mathbf{x}, \mathbf{y}) = f(\mathbf{x})^\top g(\mathbf{y}) \;\middle|\; f \in \mathcal{F}_{\texttt{NN}}(d_1, L_1, N_1, D), g \in \mathcal{F}_{\texttt{NN}}(d_2, L_2, N_2, D) \right\}. \end{aligned} \tag{12}$$

Universal approximation. Classical universal approximation theorems show that neural networks can approximate broad classes of functions [Barron, 1993]. However, our scoring class has an additional structural constraint: the score must take the inner-product form $s(\mathbf{x}, \mathbf{y}) = f(\mathbf{x})^\top g(\mathbf{y})$. It is therefore not immediate that DNN-based inner-product scores can approximate the PMI target of contrastive learning. We first show that this is indeed possible under only a mild integrability condition.

Assumption 5.1. $\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p}\left|\log \frac{p(\mathbf{x},\mathbf{y})}{p_{\mathcal{X}}(\mathbf{x}) p_{\mathcal{Y}}(\mathbf{y})}\right| < \infty$.

Theorem 5.2. *If Assumption 5.1 holds, then for any $L_1 \in \mathbb{N} \cap (1,\infty)$, $L_2 \in \mathbb{N} \cap (1,\infty)$, $\epsilon \in (0,\infty)$, there exist $N_1 \in \mathbb{N}$, $N_2 \in \mathbb{N}$, $D \in \mathbb{N}$, and $\tilde{s} \in \mathcal{F}_{\texttt{S}}(L_1, N_1, L_2, N_2, D)$ such that $\mathcal{L}(\tilde{s}) - \mathcal{L}^* \le \epsilon$.*

The proof of Theorem 5.2 is provided in Appendix C. Assumption 5.1 is very mild: it is equivalent to the finiteness of the optimal contrastive risk, i.e., $\mathcal{L}^* > -\infty$ (see Lemma C.1 for the proof of this claim), and is expected to hold in standard CRL settings. Therefore, Theorem 5.2 gives a distribution-free guarantee without any smoothness, boundedness, or low-rank assumptions, which helps explain effectiveness of DNNs across diverse datasets.

Approximation rates. The universal approximation result above is qualitative. To obtain explicit approximation rates, we impose some regularity assumption on the optimal scoring function $s^*(\mathbf{x}, \mathbf{y}) = \tau \log \frac{p(\mathbf{x},\mathbf{y})}{p_{\mathcal{X}}(\mathbf{x}) p_{\mathcal{Y}}(\mathbf{y})}$. In this work, we consider the Hölder continuity assumption.

For $\beta \in (0,1]$, $k \in \mathbb{N}$, $\Omega \subset \mathbb{R}^k$, and a real-valued function h whose domain contains Ω, the Hölder-β norm of h on Ω is defined as

$$\|h\|_{\beta,\Omega} := \max\left\{ \sup_{\mathbf{z}\in\Omega} |h(\mathbf{z})|, \sup_{\Omega \ni \mathbf{x} \ne \mathbf{y} \in \Omega} \frac{|h(\mathbf{x}) - h(\mathbf{y})|}{\|\mathbf{x} - \mathbf{y}\|_\infty^\beta} \right\}.$$

Hölder continuity is a common regularity assumption in the theoretical analysis of approximation errors (e.g. [Lu et al., 2021, Shen et al., 2022]) and generalization errors (e.g., [Schmidt-Hieber, 2020, Kim et al., 2021, Zhang et al., 2024]).

Assumption 5.3. $\mathcal{X} \times \mathcal{Y} \subset [-R, R]^{d_1+d_2}$ and $s^*(\mathbf{x}, \mathbf{y}) = \tau \log \frac{p(\mathbf{x},\mathbf{y})}{p_{\mathcal{X}}(\mathbf{x}) p_{\mathcal{Y}}(\mathbf{y})}$ satisfies $\|s^*\|_{\beta, \mathcal{X}\times\mathcal{Y}} \le B$ for some $R \in (0,\infty)$, $\beta \in (0,1]$, and $B \in (0,\infty)$.

Note that Assumption 5.3 is stronger than Assumption 5.1.

Theorem 5.4. *Let Assumption 5.3 hold. Then for a sufficiently small number $\epsilon \in (0, +\infty)$ and $D \asymp \epsilon^{-\frac{\min\{d_1,d_2\}}{2\beta}}$, $L_1 \asymp \epsilon^{-\frac{d_1}{d_2}\cdot\frac{\min\{d_1,d_2\}}{8\beta}}$, $N_1 \asymp \epsilon^{-\left(4+\frac{d_1}{d_2}\right)\cdot\frac{\min\{d_1,d_2\}}{8\beta}}$, $L_2 \asymp \epsilon^{-\frac{d_2}{d_1}\cdot\frac{\min\{d_1,d_2\}}{8\beta}}$, $N_2 \asymp \epsilon^{-\left(4+\frac{d_2}{d_1}\right)\cdot\frac{\min\{d_1,d_2\}}{8\beta}}$, it holds that $\inf\{\mathcal{L}(\tilde{s}) - \mathcal{L}^* \mid \tilde{s} \in \mathcal{F}_{\texttt{S}}(L_1, N_1, L_2, N_2, D)\} \lesssim \epsilon$.*

The proof of Theorem 5.4 is provided in Appendix D. The proof proceeds as follows. We first derive a more general result (see Theorem D.6), where we establish approximation rate for $L_1 \asymp \epsilon^{-\theta_{1,1}}$, $N_1 \asymp \epsilon^{-\theta_{1,2}}$, $L_2 \asymp \epsilon^{-\theta_{2,1}}$, $N_2 \asymp \epsilon^{-\theta_{2,2}}$, $D \asymp \epsilon^{-\theta_0}$ with $\theta_{1,1}, \theta_{1,2}, \theta_{2,1}, \theta_{2,2}, \theta_0$ being arbitrary fixed positive numbers. We then select specific $\theta_{1,1}, \theta_{1,2}, \theta_{2,1}, \theta_{2,2}, \theta_0$ to minimize the pseudo-dimension of the space $\mathcal{F}_{\mathsf{s}}(L_1, N_1, L_2, N_2, D)$ required to achieve the same approximation error $O(\epsilon)$, and Theorem D.6 with this specific choice of $\theta_{1,1}, \theta_{1,2}, \theta_{2,1}, \theta_{2,2}, \theta_0$ leads to Theorem 5.4 immediately. It turns out that the minimum pseudo-dimension is $\widetilde{O}\left(\epsilon^{-(\min\{d_1,d_2\}+d_1+d_2)/(2\beta)}\right)$, attained by the space $\mathcal{F}_{\mathsf{s}}(L_1, N_1, L_2, N_2, D)$ specified in Theorem 5.4. Recall that pseudo-dimension is widely used in learning theory to quantify the complexity of a hypothesis space (cf. Anthony and Bartlett [1999]). In Theorem G.1, we show that the pseudo-dimension of our hypothesis space $\mathcal{F}_{\mathsf{s}}(L_1, N_1, L_2, N_2, D)$ is $\widetilde{O}(L_1^2 N_1^2 + L_2^2 N_2^2 + DN_1 + DN_2)$. We use the classical result on the approximation of Hölder continuous functions by neural networks given in Shen et al. [2020] (see also Lemma D.1) as a key lemma in the proofs of Theorem D.6 and, as we shall see, Theorem 5.7.

However, relying solely on this result in Shen et al. [2020] is far from sufficient to prove these theorems. It only gives rates for approximating Hölder continuous functions by a single neural network, while in our setting, we need to approximate $s^*(\mathbf{x}, \mathbf{y}) = \tau \log \frac{p(\mathbf{x},\mathbf{y})}{p_{\mathcal{X}}(\mathbf{x}) p_{\mathcal{Y}}(\mathbf{y})}$ by the inner product of two ReLU networks $\tilde{f}(\mathbf{x})^\top \tilde{g}(\mathbf{y}) = \sum_{i=1}^D \tilde{f}_i(\mathbf{x}) \tilde{g}_i(\mathbf{y})$, rather than by a single network $\tilde{h}(\mathbf{x}, \mathbf{y})$. In fact, our proof requires a *nontrivial* finite-rank approximation $s^*(\mathbf{x}, \mathbf{y}) \approx \sum_{i=1}^D f_i(\mathbf{x}) g_i(\mathbf{y})$, where f_i (or g_i) is very easy to be approximated by ReLU networks and g_i (or f_i) is Hölder continuous. Then the only place where we invoke Shen et al. [2020] is to construct a ReLU network $\tilde{g}_i$ (or $\tilde{f}_i$) to approximate g_i (or f_i). We also emphasize the importance of Lemma D.2, which is crucial to the proof of Theorem D.6 and is established through a delicate application of Hoeffding's lemma. Lemma D.2 fundamentally bridges the uniform approximation of s^* to the risk $\mathcal{L}(\cdot) - \mathcal{L}^*$, regardless of the underlying hypothesis space, and therefore applies far beyond neural networks, thus constituting one of our major contributions.

Theorem 5.4 gives a quantitative guarantee for DNN-based inner-product CRL: it specifies how the size of networks scale with the target error ϵ. The dependence on β shows that smoother PMI targets require smaller networks, consistent with the spectral approximation result in Section 4. The dependence on d_1, d_2 reflects the standard curse of dimensionality in approximation [Bach, 2017].

A key implication of the bound is that the two encoders are *not independent*. The size of the $\mathcal{X}$-encoder depends not only on d_1, but also on d_2, and vice versa for the $\mathcal{Y}$-encoder. This is because the two networks are coupled through the shared inner-product score $f(\mathbf{x})^\top g(\mathbf{y})$, which aims to approximate the PMI function on $\mathcal{X} \times \mathcal{Y}$. Furthermore, the required dimension for the representation D only depends on $\min\{d_1, d_2\}$. Therefore, each encoder need be designed relative to the other modality in order to capture mutual information.

5.2 Approximation via DNNs: Structured Data

The previous subsection gives a general approximation for DNN-based inner-product CRL. However, the required representation dimension D can be exponentially large in the worst case. Practical data often exhibit low-dimensional semantic structure, and embeddings can capture them with moderate dimensions (e.g., 512, 1024). We next introduce a latent-class generative model (LCGM) as a tractable example to show that DNNs approximate more efficiently on structured data. In LCGM, the PMI operator is low-rank, and its singular functions admit explicit probabilistic interpretations in terms of posterior probabilities and log-posterior ratios of the latent classes.

Assumption 5.5 (Latent-class generative model, LCGM)**.** There exists a latent semantic variable $z \in [K]$ with prior $\pi_k = p(z = k)$ such that, conditional on z, the two views are independent:

$$p(\mathbf{x}, \mathbf{y}|z = k) = p(\mathbf{x}|z = k)p(\mathbf{y}|z = k). \tag{13}$$

Moreover, the anchor view is low-noise: for each $\mathbf{x} \in \mathcal{X}$, there exists a ground-truth latent label $z_{\mathbf{x}} \in [K]$ such that

$$p(z = z_{\mathbf{x}}|\mathbf{x}) = 1. \tag{14}$$

The conditional independence condition (13) is standard in theoretical analyses of contrastive learning [Chen et al., 2024, Oko et al., 2025]. The deterministic condition (14) means that the anchor view $\mathbf{x}$ reveals the latent class. Similar low-noise assumptions have been used in HaoChen et al. [2021]. This

	D	$L_1 \cdot N_1$	$L_2 \cdot N_2$	$\widetilde{\mathtt{PD}}$
General case	$\epsilon^{-\frac{\min\{d_1,d_2\}}{2\beta}}$	$\epsilon^{-(4+\frac{2d_1}{d_2})\cdot\frac{\min\{d_1,d_2\}}{8\beta}}$	$\epsilon^{-(4+\frac{2d_2}{d_1})\cdot\frac{\min\{d_1,d_2\}}{8\beta}}$	$\epsilon^{-\frac{\min\{d_1,d_2\}+d_1+d_2}{2\beta}}$
LCGM	K	$\log^2\frac{1}{\epsilon}$	$\epsilon^{-\frac{d_2}{4\beta}}$	$\epsilon^{-\frac{d_2}{2\beta}}$

Table 1: Comparison of the network sizes to achieve approximation error ϵ. Here D is the dimension of the representation, L_1, N_1 denote the depth and width of the encoder f, L_2, N_2 denote the depth and width of the encoder g, $\widetilde{\mathtt{PD}}$ is the pseudo-dimension of the space $\mathcal{F}_{\mathtt{S}}(L_1, N_1, L_2, N_2, D)$ (omitting logarithmic factors), K is the number of latent classes.

condition can be relaxed to a high-probability posterior condition at the cost of additional technical details, please refer to Appendix E.1 for more details; we keep the deterministic version for clarity. For example, z may represent the semantic class "cat", $\mathbf{x}$ a clean image of a cat, and $\mathbf{y}$ a textual description of a cat (may be noisy).

The following proposition characterizes the spectral structure of the PMI operator under LCGM. The formal statement and proof are provided in Appendix E.

Proposition 5.6 (Spectral structure under LCGM, informal of Proposition E.1)**.** *Under Assumption 5.5, there holds*

1. *The PMI operator $T : L_2(P_{\mathcal{X}}) \to L_2(P_{\mathcal{Y}})$ has rank at most K.*
2. $\operatorname{span}\{\psi_1, \ldots, \psi_K\} = \operatorname{span}\{p(z=k|\mathbf{x})\}_{k=1}^K, \operatorname{span}\{\phi_1, \ldots, \phi_K\} = \operatorname{span}\{\log p(z=k|\mathbf{y})\}_{k=1}^K$.
3. $\log \frac{p(\mathbf{x},\mathbf{y})}{p_{\mathcal{X}}(\mathbf{x})p_{\mathcal{Y}}(\mathbf{y})} = \sum_{k=1}^K p(z=k|\mathbf{x}) \log \frac{p(z=k|\mathbf{y})}{\pi_k}$

Therefore, PMI can be approximated effectively through K-dimensional representations by constructing DNNs f, g such that

$$f(\mathbf{x}) \approx [p(z=1|\mathbf{x}), \cdots, p(z=K|\mathbf{x})]^\top, \quad g(\mathbf{y}) \approx \left[\log \frac{p(z=1|\mathbf{y})}{\pi_1}, \cdots, \log \frac{p(z=K|\mathbf{y})}{\pi_K}\right]^\top.$$

In this way, the two representations have different meanings: The encoder for the low-noise modality $\mathcal{X}$ implements the classifier through learning the posterior probability $p(z=k|\mathbf{x})$, while g learns to fit the log-conditional-probability ratios $\log \frac{p(z=k|\mathbf{y})}{\pi_k}$ for each class k. Numerical experiments in Appendix I illustrate this mechanism in the synthetic LCGM setting: CRL recovers the predicted singular functions, and the effective representation dimension matches the number of latent classes.

The approximation error of DNNs for LCGM is established as follows:

Theorem 5.7. *Let Assumption 5.3 and Assumption 5.5 hold. Then for a sufficiently small number $\epsilon \in (0, \infty)$ and $L_1 \asymp \log \frac{1}{\epsilon}$, $N_1 \asymp \log \frac{1}{\epsilon}$, $L_2 \gtrsim \log \frac{1}{\epsilon}$, $N_2 \gtrsim \log \frac{1}{\epsilon}$ with $L_2 \cdot N_2 \asymp \epsilon^{-\frac{d_2}{4\beta}}$, it holds that*

$$\inf\{\mathcal{L}(\tilde{s}) - \mathcal{L}^* \mid \tilde{s} \in \mathcal{F}_{\mathtt{S}}(L_1, N_1, L_2, N_2, K)\} \lesssim \epsilon.$$

The proof of Theorem 5.7 is provided in Appendix F. Table 1 compares Theorem 5.7 with the general DNN approximation result in Theorem 5.4. The structured model leads to a substantially smaller representation and network size. In particular, the dimension of the encoder reduces from exponential term to a constant, while the depth and width of f are scaled from exponential to logarithmic terms. One can also check that $\left(4 + \frac{2d_2}{d_1}\right) \cdot \frac{\min\{d_1,d_2\}}{8\beta} > \frac{d_2}{4\beta}$. Therefore, DNNs can learn PMI efficiently on structured data by approximating the spaces spanned by the singular functions $\{\psi_k\}\{\phi_k\}$.

Remark 5.8. Although the results in this section are established for neural networks with ReLU activation, our overall proof framework *does not depend on ReLU*. If networks with another activation function satisfy the approximation properties required in our proofs, then Theorem 5.2, Theorem 5.4, and Theorem 5.7 can be extended accordingly. For example, Theorem 5.2 also holds for sigmoid networks because, according to van Nuland [2024], sigmoid networks can uniformly approximate any compactly supported continuous function on $\mathbb{R}^d$, as ReLU networks do.

6 Fast Calibration Rates under a Tsybakov Noise Condition

Now we have shown that CRL can learn representations that approximate PMI function and achieve small approximation error. In this section, we study how these representations perform in downstream retrieval. The core of the analysis is the calibration inequality which compares excess risks of the upstream and downstream tasks, such as (8). However, it suffers from low convergence rate.

In statistical learning theory, the rates can be improved under Tsybakov low-noise condition. In binary classification, the classical Tsybakov noisy condition assumes that the posterior rarely lies close to the decision boundary, which lead to faster calibration rates [Tsybakov, 2004, Zhang, 2004]. According to Lemma 3.3 of Li et al. [2026], a scoring function s achieves the optimal retrieval performance, i.e., $\mathcal{E}(s) = \mathcal{E}^*$, if and only if for all $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y}, \mathbf{y}' \in \mathcal{Y}$,

$$\left(\frac{p(\mathbf{y} \mid \mathbf{x})}{p_{\mathcal{Y}}(\mathbf{y})} - \frac{p(\mathbf{y}' \mid \mathbf{x})}{p_{\mathcal{Y}}(\mathbf{y}')}\right)\big(s(\mathbf{x}, \mathbf{y}) - s(\mathbf{x}, \mathbf{y}')\big) > 0.$$

Thus, the Bayes ranking is determined by the density ratio $\frac{p(\mathbf{y}|\mathbf{x})}{p_{\mathcal{Y}}(\mathbf{y})}$. Intuitively, the retrieval problem is easier when the density-ratio scores of candidate items are well separated. Motivated by this observation, we impose the following low-noise condition for retrieval.

Assumption 6.1 (Tsybakov low-noise condition for retrieval)**.** There exist constants $\mu \geq 0$ and $C > 0$ such that, for all $t > 0$,

$$\mathbb{P}\Big(\Big|\frac{p(\mathbf{y} \mid \mathbf{x})}{p_{\mathcal{Y}}(\mathbf{y})} - \frac{p(\mathbf{y}'|\mathbf{x})}{p_{\mathcal{Y}}(\mathbf{y}')}\Big| \leq t\Big) \leq ct^{\mu},$$

where $\mathbf{x} \sim p_{\mathbf{x}}$ and $\mathbf{y}, \mathbf{y}' \sim p_{\mathcal{Y}}$ are sampled independently. Here, μ refers to the **noise exponent.**

The above assumption always holds with $\mu = 0$. A larger exponent μ means that near-ties between candidate items occur less frequently, so the Bayes ranking has a larger margin and the retrieval problem is less noisy. Under this condition, contrastive excess risk transfers to retrieval performance at a faster rate.

Theorem 6.2 (Fast calibration under density-ratio margin)**.** *Let Assumptions 4.2 and 6.1 hold. Then*

$$\mathcal{E}^* - \mathcal{E}(s) \lesssim \big(\mathcal{L}(s) - \mathcal{L}^*\big)^{\frac{\mu+1}{\mu+2}}.$$

The proof is provided in Appendix H. Since $\frac{\mu+1}{\mu+2} \geq \frac{1}{2}$, Theorem 6.2 improves the distribution-free calibration rate in (8). Moreover, the exponent $(\mu+1)/(\mu+2)$ increases with μ and approaches 1 as $\mu \to \infty$. Therefore, when the data distribution contains fewer near-tie candidate pairs, small contrastive excess risk translates more directly into strong downstream retrieval performance.

7 Conclusion

In this paper, we developed an approximation theory for contrastive representation learning. We showed that finite-dimensional inner-product representations perform low-rank spectral approximation of the PMI operator, and that the resulting approximation error is governed by the representation dimension and the spectral decay of this operator. We then studied the realisability of such approximations by DNN encoders, establishing both a general approximation guarantee and sharper rates under structured latent-class distributions. Finally, we introduced a density-ratio Tsybakov condition under which contrastive excess risk enjoys faster calibration to downstream retrieval performance.

There are several promising future directions. It would be very interesting to study the training dynamics of CRL, for example, can first-order optimization methods learn PMI efficiently? Another problem is to study the expressive power of DNN-based cosine similarity function, which is more difficult to analyze.

Acknowledgement

The authors thank the reviewers and Area Chair for their helpful feedback and suggestions. The work is supported by Australian Research Council Discovery Project (DP250101359).

References

Ruize An, Chen Zhang, and Dawei Song. Sparse contrastive learning of sentence embeddings. *arXiv preprint arXiv:2311.03881*, 2023.

Martin Anthony and Peter L Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, New York, NY, 1999.

Jordan Ash, Surbhi Goel, Akshay Krishnamurthy, and Dipendra Misra. Investigating the role of negatives in contrastive representation learning. In *International Conference on Artificial Intelligence and Statistics*, pages 7187–7209. PMLR, 2022.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017. URL `http://jmlr.org/papers/v18/14-546.html`.

Parikshit Bansal, Ali Kavis, and Sujay Sanghavi. Understanding self-supervised learning via Gaussian mixture models, 2026. URL `https://arxiv.org/abs/2411.03517`.

Han Bao, Yoshihiro Nagano, and Kento Nozawa. On the surrogate gap between contrastive and supervised losses. In *International conference on machine learning*, pages 1585–1606. PMLR, 2022.

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

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

Roy Betser, Eyal Gofer, Meir Yossef Levi, and Guy Gilboa. InfoNCE induces Gaussian distribution. *arXiv preprint arXiv:2602.24012*, 2026.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013. ISBN 978-0-19-953525-5.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Zixiang Chen, Yihe Deng, Yuanzhi Li, and Quanquan Gu. Understanding transferable representation learning and zero-shot transfer in clip, 2024. URL `https://arxiv.org/abs/2310.00927`.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

J. C. Ferreira and V. A. Menegatto. Eigenvalues of integral operators defined by smooth positive definite kernels. *Integral Equations and Operator Theory*, 64(1):61–81, 2009.

Gerald B Folland. *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, New York, NY, second edition, 1999.

Xin Guo, Zheng-Chu Guo, and Lei Shi. Capacity dependent analysis for functional online learning algorithms. *Applied and Computational Harmonic Analysis*, 67:101567, 2023.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in neural information processing systems*, 34:5000–5011, 2021.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning, 2022. URL https://arxiv.org/abs/2110.09348.

Daniel D. Johnson, Ayoub El Hanchi, and Chris J. Maddison. Contrastive learning can find an optimal basis for approximately view-invariant functions, 2023. URL https://arxiv.org/abs/2210.01883.

Dhruva Karkada, James Simon, Yasaman Bahri, and Michael Deweese. Closed-form training dynamics reveal learned features and linear structure in word2vec-like models. *Advances in neural information processing systems*, 38:3967–4000, 2025.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781, 2020.

Yongdai Kim, Ilsang Ohn, and Dongha Kim. Fast convergence rates of deep neural networks for classification. *Neural Networks*, 138:179–197, 2021.

Yunwen Lei, Tianbao Yang, Yiming Ying, and Ding-Xuan Zhou. Generalization analysis for contrastive representation learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19200–19227. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/lei23a.html.

Moshe Leshno, Vladimir Ya Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, 27, 2014.

Yuanfan Li, Xiyuan Wei, Tianbao Yang, and Yiming Ying. Statistical consistency and generalization of contrastive representation learning. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.02116.

Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.

E. J. McShane. Extension of range of functions. *Bulletin of the American Mathematical Society*, 40 (12):837–842, 1934.

Thomas Nagler and Sophie Langer. Optimal neural network approximation of smooth compositional functions on sets with low intrinsic dimension. In Steve Hanneke and Tor Lattimore, editors, *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pages 5244–5272. PMLR, 29 Jun–03 Jul 2026. URL https://proceedings.mlr.press/v336/nagler26a.html.

Kazusato Oko, Licong Lin, Yuhang Cai, and Song Mei. A statistical theory of contrastive pre-training and multimodal generative ai, 2025. URL https://arxiv.org/abs/2501.04641.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, New York, third edition, 1976.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5628–5637. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/saunshi19a.html.

Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 19250–19286. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/saunshi22a.html.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation characterized by number of neurons. *Communications in Computational Physics*, 28(5):1768–1811, 2020.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Optimal approximation rate of ReLU networks in terms of width and depth. *Journal de Mathématiques Pures et Appliquées*, 157:101–135, 2022. doi: 10.1016/j.matpur.2021.07.009.

Yuandong Tian, Lantao Yu, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning with dual deep networks, 2021. URL https://arxiv.org/abs/2010.00578.

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

Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Toshimitsu Uesaka, Taiji Suzuki, Yuhta Takida, Chieh-Hsin Lai, Naoki Murata, and Yuki Mitsufuji. Weighted point set embedding for multimodal contrastive learning toward optimal similarity metric, 2025. URL https://arxiv.org/abs/2404.19228.

Teun D. H. van Nuland. Noncompact uniform universal approximation. *Neural Networks*, 173: 106181, 2024.

Bokun Wang, Yunwen Lei, Yiming Ying, and Tianbao Yang. On discriminative probabilistic modeling for self-supervised representation learning, 2025. URL https://arxiv.org/abs/2410.09156.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94: 103–114, 2017.

Yiming Ying and Massimiliano Pontil. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 8(5):561–596, 2008.

Zhuoning Yuan, Yuexin Wu, Zi-Hao Qiu, Xianzhi Du, Lijun Zhang, Denny Zhou, and Tianbao Yang. Provable stochastic optimization for global contrastive learning: Small batch does not harm performance. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 25760–25782. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/yuan22b.html.

Runtian Zhai, Bingbin Liu, Andrej Risteski, J Zico Kolter, and Pradeep Kumar Ravikumar. Understanding augmentation-based self-supervised representation learning via RKHS approximation and regression. In *The Twelfth International Conference on Learning Representations*, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11941–11952. IEEE, 2023.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.

Yuhui Zhang, Yuichiro Wada, Hiroki Waida, Kaito Goto, Yusaku Hino, and Takafumi Kanamori. Deep clustering with a constraint for topological invariance based on symmetric InfoNCE. *Neural Computation*, 35(7):1288–1339, 2023.

Zihan Zhang, Lei Shi, and Ding-Xuan Zhou. Classification with deep neural networks and logistic loss. *Journal of Machine Learning Research*, 25(125):1–117, 2024.

Ding-Xuan Zhou. Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, 48(2):787–794, 2020.

A Further Related Work

InfoNCE loss. InfoNCE is one of the most widely used objectives in contrastive representation learning. It contrasts each anchor with a finite number of negative samples, usually drawn from the same mini-batch, and has achieved strong empirical success in self-supervised and multimodal learning [Oord et al., 2018, Chen et al., 2020, Radford et al., 2021]. InfoNCE has also been connected to mutual-information maximization [Oord et al., 2018, Zhang et al., 2023]. In practice, however, algorithms based on InfoNCE often benefit from large batch sizes; for example, CLIP uses very large batches during training [Radford et al., 2021]. To reduce this dependence, Yuan et al. [2022] proposed the global contrastive loss (GCL), which contrasts each anchor against all negatives , and developed SogCLR as an efficient stochastic optimization method that does not rely on large batch sizes. There are also some works using different losses for contrastive representation learning, for example, SigLIP[Zhai et al., 2023], which replaces batch-wise softmax by pairwise sigmoid losses for matched and unmatched pairs. At the population level, this is binary classification between positive pairs from joint distribution p and negative pairs from $P_X \otimes P_Y$. Moreover, the global minimizer of logistic loss is log-density-ratio $\log \frac{p(x,y)}{p_X(x)p_Y(y)}$. Therefore, the optimal scoring function for SigLIP is exactly PMI, making our approximation analysis relevant to SigLIP.

Other theories of contrastive representation learning. Several other theoretical perspectives have been proposed to understand CRL. Wang and Isola [2020] analyzed contrastive learning through the alignment and uniformity properties of representations on the hypersphere. Other works study the training dynamics of contrastive objectives, including dimensional collapse and feature learning dynamics [Jing et al., 2022, Tian et al., 2021]. Wang et al. [2025] analyzed self-supervised contrastive learning from the perspective of discriminative probabilistic modeling, while Tsai et al. [2020] studied self-supervised learning under a multi-view framework. Some works study InfoNCE from the perspective of Word2Vec and its training dynamics [Levy and Goldberg, 2014, Karkada et al., 2025]. More recently, Bansal et al. [2026] investigated contrastive learning under Gaussian mixture models and showed that contrastive learning can recover low-dimensional structure. Some work also shows that InfoNCE induces Gaussian distribution [Betser et al., 2026].

Approximation via DNNs. Deep neural networks have shown to admit universal approximation ability [Barron, 1993]. Error bounds for deep ReLU neural networks and CNNs have been established [Yarotsky, 2017, Zhou, 2020]. Following the idea of Yarotsky [2017], it is proved in Schmidt-Hieber [2020] that

$$\sup_{\substack{h:[-R,R]^d\to\mathbb{R}\\ \|h\|_{\beta,[-R,R]^d}}} \inf \left\{ \sup_{[-R,R]^d} |f-h| \;\middle|\; \begin{array}{l} f \in \mathcal{F}_{\texttt{NN}}\left(d, \log\frac{1}{\epsilon}, \epsilon^{-\frac{d}{\beta}}, 1\right) \text{ and the number} \\ \text{of nonzero parameters of } f \lesssim \epsilon^{-\frac{d}{\beta}} \cdot \log\frac{1}{\epsilon} \end{array} \right\} \lesssim \epsilon.$$

This rate turns to be sharp, and many results use this approximation rate to obtain optimal (up to logarithmic terms) convergence rates for generalization errors in various machine learning problems, such as regression [Schmidt-Hieber, 2020] and classification [Kim et al., 2021, Zhang et al., 2024]. However, the restriction

$$\text{the number of nonzero parameters of } f \lesssim \epsilon^{-\frac{d}{\beta}} \cdot \log\frac{1}{\epsilon} \tag{15}$$

makes training over the hypothesis space

$$\left\{ f \;\middle|\; \begin{array}{l} f \in \mathcal{F}_{\texttt{NN}}\left(d, \log\frac{1}{\epsilon}, \epsilon^{-\frac{d}{\beta}}, 1\right) \text{ and the number} \\ \text{of nonzero parameters of } f \lesssim \epsilon^{-\frac{d}{\beta}} \cdot \log\frac{1}{\epsilon} \end{array} \right\}$$

impossible to implement. In Shen et al. [2020, 2022], Lu et al. [2021], the authors use a novel technique based on bit extraction to obtain sharp neural network approximation rates with the restriction like (15) removed. Specifically, they establish bounds of the form

$$\sup_{\substack{h:[-R,R]^d\to\mathbb{R}\\ \|h\|_{\beta,[-R,R]^d}}} \inf \left\{ \sup_{[-R,R]^d} |f-h| \;\middle|\; f \in \mathcal{F}_{\texttt{NN}}(d, L, N, 1) \right\} \lesssim \left(\frac{\log^{\iota_1} L}{L} \cdot \frac{\log^{\iota_2} N}{N} \right)^{2\beta/d}, \tag{16}$$

where $\iota_1, \iota_2 \in \mathbb{R}$ are absolute constants. In the recent work [Nagler and Langer, 2026], the authors obtain (16) with the additional requirement that all network weights and biases are polynomially bounded in N. Such control over parameter magnitudes may be useful in other settings. In this work, we will mainly use Theorem 1.1 of Shen et al. [2020] to construct neural networks $\tilde{f}(\mathbf{x})$ and $\tilde{g}(\mathbf{y})$, such that $\tilde{f}(\mathbf{x})^\top \tilde{g}(\mathbf{y})$ can approximate the Hölder-β continuous target function $s^*(\mathbf{x}, \mathbf{y}) = \tau \log \frac{p(\mathbf{x},\mathbf{y})}{p_\mathcal{X}(\mathbf{x})p_\mathcal{Y}(\mathbf{y})}$ well.

B Proofs in Section 4

B.1 Proof of Lemma 4.1

Proof of Lemma 4.1. For $\lambda_i, \lambda_j > 0$, we have

$$\begin{aligned}\langle \lambda_i^{-1/2} T\psi_i, \lambda_j^{-1/2} T\psi_j \rangle_{P_\mathcal{Y}} &= \langle \lambda_i^{-1/2} T^* T\psi_i, \lambda_j^{-1/2} \psi_j \rangle_{P_\mathcal{X}} \\ &= \frac{\sqrt{\lambda_i}}{\sqrt{\lambda_j}} \delta_{ij}.\end{aligned}$$

where the first equality is because T, T^* are adjoint. Therefore, we can let $\phi_i = \lambda_i^{-1/2} T\psi_i$ and then $T\psi_i = \sqrt{\lambda_i}\phi_i$. The inverse direction also holds similarly.

For almost every $\mathbf{x}$, $\log \frac{p(\mathbf{x},\mathbf{y})}{p_\mathcal{X}(\mathbf{x})p_\mathcal{Y}(\mathbf{y})} \in L_2(P_\mathcal{Y})$. We have

$$\left\langle \log \frac{p(\mathbf{x},\mathbf{y})}{p_\mathcal{X}(\mathbf{x})p_\mathcal{Y}(\mathbf{y})}, \phi_i \right\rangle_{P_\mathcal{Y}} = \int \log \frac{p(\mathbf{x},\mathbf{y})}{p_\mathcal{X}(\mathbf{x})p_\mathcal{Y}(\mathbf{y})} \phi_i(\mathbf{y}) p_\mathcal{Y}(\mathbf{y}) d\mathbf{y} = (T^*\phi_i)(\mathbf{x}) = \sqrt{\lambda_i}\psi_i(\mathbf{x}).$$

Therefore, $\log \frac{p(\mathbf{x},\mathbf{y})}{p_\mathcal{X}(\mathbf{x})p_\mathcal{Y}(\mathbf{y})} = \sum_{i=1}^\infty \sqrt{\lambda_i}\psi_i(\mathbf{x})\phi_i(\mathbf{y})$.

Now consider $g \in L_2(P_\mathcal{Y})$ with coefficients $a_i = \langle g, \phi_i \rangle$. Then $T^* g = \sum_i a_i \sqrt{\lambda_i}\psi_i$, and

$$\mathrm{Im}(T^*) = \left\{ \sum_{i:\lambda_i>0} c_i\psi_i : \sum_i c_i^2/\lambda_i < \infty \right\} = \mathcal{H}_\mathcal{X}, \qquad \left\| \sum_i c_i\psi_i \right\|_{\mathcal{H}_\mathcal{X}}^2 = \sum_i c_i^2/\lambda_i.$$

This is exactly the RKHS of $K_\mathcal{X}$ by its feature-map construction. Interchanging $\mathcal{X}$ and $\mathcal{Y}$ gives the corresponding claim for T and $\mathcal{H}_\mathcal{Y}$. □

B.2 Proof of Theorem 4.3

To start, we provide the following lemma, which bounds the excess risk of contrastive loss through L_2 distance.

Lemma B.1. *Let Assumption 4.2 hold, let $s^* = \tau \log r \in L_2(P_\mathcal{X} \otimes P_\mathcal{Y})$, where $r(\mathbf{x}, \mathbf{y}) = p(\mathbf{y} \mid \mathbf{x})/p_\mathcal{Y}(\mathbf{y})$, and let $\|s\|_\infty \le B$. Then*

$$\mathcal{L}(s) - \mathcal{L}(s^*) \le \frac{\alpha e^{2B/\tau}}{2\tau} \|s - s^*\|_{L_2(P_\mathcal{X} \otimes P_\mathcal{Y})}^2.$$

Proof. Denote $\Delta := s - s^*$ and, for $\theta \in [0, 1]$, define $s_\theta := s^* + \theta\Delta$ and $G(\theta) := \mathcal{L}(s_\theta)$. Let

$$Z_\theta(\mathbf{x}) := \mathbb{E}_{\mathbf{y}\sim p_\mathcal{Y}} \exp\left(\frac{s_\theta(\mathbf{x},\mathbf{y})}{\tau}\right), \qquad q_{\theta,\mathbf{x}}(\mathbf{y}) := \frac{p_\mathcal{Y}(\mathbf{y}) \exp(s_\theta(\mathbf{x},\mathbf{y})/\tau)}{Z_\theta(\mathbf{x})}.$$

Thus $q_{\theta,\mathbf{x}}$ is a probability density on $\mathcal{Y}$ induced by the scoring function $s_\theta(\mathbf{x}, \cdot)$. Since $r \le \alpha$ and $\mathbb{E}_{\mathbf{y}\sim p_\mathcal{Y}} r(\mathbf{x}, \mathbf{y}) = \int p_\mathcal{Y}(\mathbf{y}) \frac{p(\mathbf{y}|\mathbf{x})}{p_\mathcal{Y}(\mathbf{y})} d\mathbf{y} = \int p(\mathbf{y}|\mathbf{x}) d\mathbf{y} = 1$, we have

$$\mathbb{E}_{\mathbf{y}\sim p_\mathcal{Y}} r(\mathbf{x}, \mathbf{y})^{1-\theta} \ge \alpha^{-\theta} \mathbb{E}_{\mathbf{y}\sim p_\mathcal{Y}} r(\mathbf{x}, \mathbf{y}) = \alpha^{-\theta}.$$

By the definition of the contrastive risk,

$$G(\theta) = \tau \mathbb{E}_\mathbf{x} \log Z_\theta(\mathbf{x}) - \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p} s_\theta(\mathbf{x}, \mathbf{y}).$$

Differentiating along this path gives

$$G'(\theta) = \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbf{y}\sim q_{\theta,\mathbf{x}}}\Delta(\mathbf{x},\mathbf{y}) - \mathbb{E}_{\mathbf{y}\sim p(\cdot|\mathbf{x})}\Delta(\mathbf{x},\mathbf{y})\right],$$
$$G''(\theta) = \frac{1}{\tau}\mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbf{y}\sim q_{\theta,\mathbf{x}}}\Delta(\mathbf{x},\mathbf{y})^2 - \left(\mathbb{E}_{\mathbf{y}\sim q_{\theta,\mathbf{x}}}\Delta(\mathbf{x},\mathbf{y})\right)^2\right]$$
$$= \frac{1}{\tau}\mathbb{E}_{\mathbf{x}}\operatorname{Var}_{\mathbf{y}\sim q_{\theta,\mathbf{x}}}\big(\Delta(\mathbf{x},\mathbf{y})\big).$$

At $\theta = 0$, $Z_0(\mathbf{x}) = 1$ and $q_{0,\mathbf{x}}(\mathbf{y}) = p(\mathbf{y} \mid \mathbf{x})$, so $G'(0) = 0$. The second-order Taylor formula with integral remainder therefore yields

$$\mathcal{L}(s) - \mathcal{L}(s^*) = G(1) - G(0) = G'(0) + \int_0^1 (1-\theta)G''(\theta)\,d\theta$$
$$= \frac{1}{\tau}\int_0^1 (1-\theta)\mathbb{E}_{\mathbf{x}}\operatorname{Var}_{\mathbf{y}\sim q_{\theta,\mathbf{x}}}\big(\Delta(\mathbf{x},\mathbf{y})\big)\,d\theta. \tag{17}$$

We now bound the variance in (17). The lower bound on Z_θ and $|s| \le B$ imply

$$\frac{q_{\theta,\mathbf{x}}(\mathbf{y})}{p_{\mathcal{Y}}(\mathbf{y})} = \frac{r(\mathbf{x},\mathbf{y})^{1-\theta}e^{\theta s(\mathbf{x},\mathbf{y})/\tau}}{Z_\theta(\mathbf{x})} \le \frac{\alpha^{1-\theta}e^{\theta B/\tau}}{\alpha^{-\theta}e^{-\theta B/\tau}} = \alpha e^{2\theta B/\tau} \le \alpha e^{2B/\tau}.$$

Consequently,

$$\operatorname{Var}_{\mathbf{y}\sim q_{\theta,\mathbf{x}}}\big(\Delta(\mathbf{x},\mathbf{y})\big) = \inf_{c\in\mathbb{R}}\mathbb{E}_{\mathbf{y}\sim q_{\theta,\mathbf{x}}}\left[(\Delta(\mathbf{x},\mathbf{y}) - c)^2\right]$$
$$\le \alpha e^{2B/\tau}\inf_{c\in\mathbb{R}}\mathbb{E}_{\mathbf{y}\sim p_{\mathcal{Y}}}\left[(\Delta(\mathbf{x},\mathbf{y}) - c)^2\right]$$
$$= \alpha e^{2B/\tau}\operatorname{Var}_{\mathbf{y}\sim p_{\mathcal{Y}}}\big(\Delta(\mathbf{x},\mathbf{y})\big).$$

Substituting this variance bound into (17) gives

$$\mathcal{L}(s) - \mathcal{L}(s^*) \le \frac{\alpha e^{2B/\tau}}{\tau}\int_0^1 (1-\theta)\,d\theta\,\mathbb{E}_{\mathbf{x}}\operatorname{Var}_{\mathbf{y}\sim p_{\mathcal{Y}}}\big(\Delta(\mathbf{x},\mathbf{y})\big)$$
$$\le \frac{\alpha e^{2B/\tau}}{2\tau}\mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}\sim p_{\mathcal{Y}}}\Delta(\mathbf{x},\mathbf{y})^2$$
$$= \frac{\alpha e^{2B/\tau}}{2\tau}\|s - s^*\|^2_{L_2(P_{\mathcal{X}}\otimes P_{\mathcal{Y}})}.$$

This completes the proof. □

Proof of Theorem 4.3. Use $s_D = \tau\sum_{i=1}^D \sqrt{\lambda_i}\psi_i\phi_i \in \mathcal{S}_D$. Orthogonality gives $\|s_D - s^*\|_2^2 = \tau^2\sum_{i>D}\lambda_i$. The uniform bound on s_D and Lemma B.1 imply

$$\inf_{s\in\mathcal{S}_D}(\mathcal{L}(s) - \mathcal{L}^*) \le \mathcal{L}(s_D) - \mathcal{L}^* \le \frac{\alpha\tau e^{2B/\tau}}{2}\sum_{i>D}\lambda_i.$$

The proof is completed. □

B.3 RKHS balls and cosine scores

Approximation analysis for RKHSs. Let $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ be the RKHSs introduced in Lemma 4.1. For D-dimensional encoders f and g, define

$$\|f\|^2_{\mathcal{H}^D_{\mathcal{X}}} = \textstyle\sum_{\ell=1}^D \|f_\ell\|^2_{\mathcal{H}_{\mathcal{X}}}, \qquad \|g\|^2_{\mathcal{H}^D_{\mathcal{Y}}} = \textstyle\sum_{\ell=1}^D \|g_\ell\|^2_{\mathcal{H}_{\mathcal{Y}}},$$

and consider

$$\mathcal{S}_{D,\kappa} = \left\{s(\mathbf{x},\mathbf{y}) = f(\mathbf{x})^\top g(\mathbf{y}) : \|f\|_{\mathcal{H}^D_{\mathcal{X}}} \le \kappa,\ \|g\|_{\mathcal{H}^D_{\mathcal{Y}}} \le \kappa\right\}.$$

Write $\mathrm{PMI}(\mathbf{x},\mathbf{y}) = \sum_{j\geq 1}\sqrt{\lambda_j}\psi_j(\mathbf{x})\phi_j(\mathbf{y})$, and set

$$f_j(\mathbf{x}) = \sqrt{\tau}\lambda_j^{1/4}\psi_j(\mathbf{x}), \qquad g_j(\mathbf{y}) = \sqrt{\tau}\lambda_j^{1/4}\phi_j(\mathbf{y}).$$

For the first r components,

$$\|f\|^2_{\mathcal{H}^r_{\mathcal{X}}} = \|g\|^2_{\mathcal{H}^r_{\mathcal{Y}}} = \tau\sum_{j=1}^{r}\lambda_j^{-1/2}.$$

Define

$$r_{D,\kappa} = \max\left\{r \leq D : \tau\sum_{j=1}^{r}\lambda_j^{-1/2} \leq \kappa^2\right\}.$$

The first $r_{D,\kappa}$ spectral components then belong to $\mathcal{S}_{D,\kappa}$, so

$$\inf_{s\in\mathcal{S}_{D,\kappa}}\|s-s^*\|^2_{L_2(P_{\mathcal{X}}\otimes P_{\mathcal{Y}})} \leq \tau^2\sum_{j>r_{D,\kappa}}\lambda_j.$$

Thus, increasing κ allows more spectral components to be represented. In particular, if

$$\kappa^2 \geq \tau\sum_{j=1}^{D}\lambda_j^{-1/2},$$

then the full rank-D spectral truncation belongs to the RKHS ball and

$$\inf_{s\in\mathcal{S}_{D,\kappa}}\|s-s^*\|^2_{L_2(P_{\mathcal{X}}\otimes P_{\mathcal{Y}})} \leq \tau^2\sum_{j>D}\lambda_j.$$

For fixed D, the bound reaches the rank-D spectral tail once κ is sufficiently large; increasing both D and κ allows the approximation error to converge to zero.

Cosine similarity. Let

$$s_D(\mathbf{x},\mathbf{y}) = \tau\sum_{j=1}^{D}\sqrt{\lambda_j}\psi_j(\mathbf{x})\phi_j(\mathbf{y}) = \tau f_D(\mathbf{x})^\top g_D(\mathbf{y})$$

be the rank-D spectral truncation of the PMI, where $f_D(\mathbf{x})_j = \lambda_j^{1/4}\psi_j(\mathbf{x}), g_D(\mathbf{y})_j = \lambda_j^{1/4}\phi_j(\mathbf{y})$. Choose $a,b,\tau > 0$ such that $\sup_{\mathbf{x},j} af_D(\mathbf{x})_j \leq 1, \sup_{\mathbf{y},j} bg_D(\mathbf{y})_j \leq 1$ and $\tau = ab$. We further define

$$u_D(\mathbf{x}) = \left(af_D(\mathbf{x}), \sqrt{1-a^2\|f_D(\mathbf{x})\|^2}, 0\right), v_D(\mathbf{y}) = \left(bg_D(\mathbf{y}), 0, \sqrt{1-b^2\|g_D(\mathbf{y})\|^2}\right).$$

Then $u_D(\mathbf{x}), v_D(\mathbf{y}) \in \mathbb{S}^{D+1}$ and $u_D(\mathbf{x})^\top v_D(\mathbf{y}) = \tau f_D(\mathbf{x})^\top g_D(\mathbf{y}) = s_D(\mathbf{x},\mathbf{y})$. Therefore, $\|s^*(\mathbf{x},\mathbf{y}) - u_D(\mathbf{x})^\top v_D(\mathbf{y})\|^2_{L_2(P_{\mathcal{X}}\otimes P_{\mathcal{Y}})} = \tau^2\sum_{j>D}\lambda_j$.

Thus, the rank-D spectral comparator can be represented exactly by a cosine-similarity model in dimension $D+2$.

C Proof of Theorem 5.2

Lemma C.1. $\mathcal{L}^* = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p}\left[\tau \cdot \log \frac{p_{\mathcal{Y}}(\mathbf{y})}{p(\mathbf{y}|\mathbf{x})}\right] < +\infty$. *Moreover,* $\mathcal{L}^* > -\infty$ *if and only if*

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p}\left[\left|\log \frac{p(\mathbf{x},\mathbf{y})}{p_{\mathcal{X}}(\mathbf{x})p_{\mathcal{Y}}(\mathbf{y})}\right|\right] < \infty.$$

Proof. Denote $s^*(\mathbf{x},\mathbf{y}) := \tau \cdot \log \frac{p(\mathbf{y}|\mathbf{x})}{p_{\mathcal{Y}}(\mathbf{y})}$. For simplicity, we denote $p^+_{\mathbf{x}} := p(\,\cdot \mid \mathbf{x})$. According to Lemma 3.1,

$$\begin{aligned}
\mathcal{L}^* = \mathcal{L}(s^*) &= \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}\sim p^+_{\mathbf{x}}}\left[\tau \log \mathbb{E}_{\mathbf{y}'\sim p^-_{\mathbf{x}}} \exp\left(\frac{s^*(\mathbf{x},\mathbf{y}') - s^*(\mathbf{x},\mathbf{y})}{\tau}\right)\right] \\
&= \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}\sim p^+_{\mathbf{x}}}\left[\tau \log\left(\mathbb{E}_{\mathbf{y}'\sim p_{\mathcal{Y}}} \frac{p^+_{\mathbf{x}}(\mathbf{y}')}{p_{\mathcal{Y}}(\mathbf{y}')} \cdot \frac{p_{\mathcal{Y}}(\mathbf{y})}{p^+_{\mathbf{x}}(\mathbf{y})}\right)\right] \\
&= \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}\sim p^+_{\mathbf{x}}}\left[\tau \log\left(\int p_{\mathcal{Y}}(\mathbf{y}') \cdot \frac{p^+_{\mathbf{x}}(\mathbf{y}')}{p_{\mathcal{Y}}(\mathbf{y}')} \cdot \frac{p_{\mathcal{Y}}(\mathbf{y})}{p^+_{\mathbf{x}}(\mathbf{y})} d\mathbf{y}'\right)\right] \\
&= \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}\sim p^+_{\mathbf{x}}}\left[\tau \log\left(\frac{p_{\mathcal{Y}}(\mathbf{y})}{p^+_{\mathbf{x}}(\mathbf{y})} \int p^+_{\mathbf{x}}(\mathbf{y}')d\mathbf{y}'\right)\right] \\
&= \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}\sim p^+_{\mathbf{x}}}\left[\tau \log \frac{p_{\mathcal{Y}}(\mathbf{y})}{p^+_{\mathbf{x}}(\mathbf{y})}\right] = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p}\left[\tau \log \frac{p_{\mathcal{Y}}(\mathbf{y})}{p(\mathbf{y} \mid \mathbf{x})}\right].
\end{aligned}$$

Note that

$$\begin{aligned}
&\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p}\left[\text{the positive part of } \log \frac{p_{\mathcal{Y}}(\mathbf{y})}{p(\mathbf{y} \mid \mathbf{x})}\right] = \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}\sim p^+_{\mathbf{x}}}\left[\max\left\{\log \frac{p_{\mathcal{Y}}(\mathbf{y})}{p^+_{\mathbf{x}}(\mathbf{y})}, 0\right\}\right] \\
&\leq \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}\sim p^+_{\mathbf{x}}}\left[\frac{p_{\mathcal{Y}}(\mathbf{y})}{p^+_{\mathbf{x}}(\mathbf{y})}\right] = \mathbb{E}_{\mathbf{x}}\left[\int \frac{p_{\mathcal{Y}}(\mathbf{y})}{p^+_{\mathbf{x}}(\mathbf{y})} \cdot p^+_{\mathbf{x}}(\mathbf{y})d\mathbf{y}\right] = \mathbb{E}_{\mathbf{x}}\left[\int p_{\mathcal{Y}}(\mathbf{y})d\mathbf{y}\right] = \mathbb{E}_{\mathbf{x}}[1] = 1 < +\infty.
\end{aligned}$$

Thus $\mathcal{L}^* = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p}\left[\tau \log \frac{p_{\mathcal{Y}}(\mathbf{y})}{p(\mathbf{y}|\mathbf{x})}\right] < +\infty$, and

$$\begin{aligned}
\mathcal{L}^* > -\infty &\iff \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p}\left[\log \frac{p_{\mathcal{Y}}(\mathbf{y})}{p(\mathbf{y} \mid \mathbf{x})}\right] > -\infty \\
&\iff \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p}\left[\text{the negative part of } \log \frac{p_{\mathcal{Y}}(\mathbf{y})}{p(\mathbf{y} \mid \mathbf{x})}\right] < \infty \\
&\iff \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p}\left[\left|\log \frac{p_{\mathcal{Y}}(\mathbf{y})}{p(\mathbf{y} \mid \mathbf{x})}\right|\right] < \infty \iff \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p}\left[\left|\log \frac{p(\mathbf{y} \mid \mathbf{x})}{p_{\mathcal{Y}}(\mathbf{y})}\right|\right] < \infty \\
&\iff \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p}\left[\left|\log \frac{p(\mathbf{x},\mathbf{y})}{p_{\mathcal{X}}(\mathbf{x})p_{\mathcal{Y}}(\mathbf{y})}\right|\right] < \infty.
\end{aligned}$$

This completes the proof. □

Lemma C.2. *Let Assumption 5.1 hold. Denote* $s^*(\mathbf{x},\mathbf{y}) := \tau \log \frac{p(\mathbf{y}|\mathbf{x})}{p_{\mathcal{Y}}(\mathbf{y})}$. *Then for any bounded measurable function* $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ *there holds*

$$\mathcal{L}(s) - \mathcal{L}^* = \mathbb{E}_{\mathbf{x}}\left[\tau \log \mathbb{E}_{\mathbf{y}\sim p(\cdot|\mathbf{x})} \exp\left((s(\mathbf{x},\mathbf{y}) - s^*(\mathbf{x},\mathbf{y}))/\tau\right)\right] - \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p}\left[s(\mathbf{x},\mathbf{y}) - s^*(\mathbf{x},\mathbf{y})\right].$$

Proof. For simplicity, we denote $p^+_{\mathbf{x}} := p(\,\cdot \mid \mathbf{x})$. It follows from Lemma C.1 that

$$\mathcal{L}^* = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p}\left[\tau \log \frac{p_{\mathcal{Y}}(\mathbf{y})}{p(\mathbf{y} \mid \mathbf{x})}\right] = -\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p}\left[\tau \log \frac{p(\mathbf{y} \mid \mathbf{x})}{p_{\mathcal{Y}}(\mathbf{y})}\right] = -\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p}\left[s^*(\mathbf{x},\mathbf{y})\right]. \quad (18)$$

We also have that

$$\begin{aligned}
&\mathcal{L}(s) = \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}\sim p_{\mathbf{x}}^+}\left[\tau \log \mathbb{E}_{\mathbf{y}'\sim p_{\mathcal{Y}}} \exp\left(\frac{s(\mathbf{x},\mathbf{y}') - s(\mathbf{x},\mathbf{y})}{\tau}\right)\right] \\
&= \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}\sim p_{\mathbf{x}}^+}\left[\tau \log \mathbb{E}_{\mathbf{y}'\sim p_{\mathcal{Y}}} \frac{\exp\left(s(\mathbf{x},\mathbf{y}')/\tau\right)}{\exp\left(s(\mathbf{x},\mathbf{y})/\tau\right)}\right] = \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}\sim p_{\mathbf{x}}^+}\left[\tau \log \frac{\mathbb{E}_{\mathbf{y}'\sim p_{\mathcal{Y}}} \exp\left(s(\mathbf{x},\mathbf{y}')/\tau\right)}{\exp\left(s(\mathbf{x},\mathbf{y})/\tau\right)}\right] \\
&= \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}\sim p_{\mathbf{x}}^+}\left[\tau \log \mathbb{E}_{\mathbf{y}'\sim p_{\mathcal{Y}}} \exp\left(s(\mathbf{x},\mathbf{y}')/\tau\right)\right] - \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}\sim p_{\mathbf{x}}^+}\left[\tau \log \exp\left(s(\mathbf{x},\mathbf{y})/\tau\right)\right] \\
&= \mathbb{E}_{\mathbf{x}}\left[\tau \log \mathbb{E}_{\mathbf{y}'\sim p_{\mathcal{Y}}} \exp\left(s(\mathbf{x},\mathbf{y}')/\tau\right)\right] - \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}\sim p_{\mathbf{x}}^+}\left[s(\mathbf{x},\mathbf{y})\right] \\
&= \mathbb{E}_{\mathbf{x}}\left[\tau \log \mathbb{E}_{\mathbf{y}'\sim p_{\mathbf{x}}^+}\left[\exp\left(s(\mathbf{x},\mathbf{y}')/\tau\right) \cdot \frac{p_{\mathcal{Y}}(\mathbf{y}')}{p_{\mathbf{x}}^+(\mathbf{y}')}\right]\right] - \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p}\left[s(\mathbf{x},\mathbf{y})\right] \\
&= \mathbb{E}_{\mathbf{x}}\left[\tau \log \mathbb{E}_{\mathbf{y}'\sim p_{\mathbf{x}}^+}\left[\exp\left(s(\mathbf{x},\mathbf{y}')/\tau\right) \cdot \exp(-s^*(\mathbf{x},\mathbf{y}')/\tau)\right]\right] - \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p}\left[s(\mathbf{x},\mathbf{y})\right] \\
&= \mathbb{E}_{\mathbf{x}}\left[\tau \log \mathbb{E}_{\mathbf{y}\sim p_{\mathbf{x}}^+}\exp\left(\left(s(\mathbf{x},\mathbf{y}) - s^*(\mathbf{x},\mathbf{y})\right)/\tau\right)\right] - \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p}\left[s(\mathbf{x},\mathbf{y})\right],
\end{aligned}$$

which, together with (18), yields

$$\begin{aligned}
&\mathcal{L}(s) - \mathcal{L}^* \\
&= \mathbb{E}_{\mathbf{x}}\left[\tau \log \mathbb{E}_{\mathbf{y}\sim p_{\mathbf{x}}^+}\exp\left(\left(s(\mathbf{x},\mathbf{y}) - s^*(\mathbf{x},\mathbf{y})\right)/\tau\right)\right] - \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p}\left[s(\mathbf{x},\mathbf{y})\right] + \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p}\left[s^*(\mathbf{x},\mathbf{y})\right] \\
&= \mathbb{E}_{\mathbf{x}}\left[\tau \log \mathbb{E}_{\mathbf{y}\sim p(\cdot|\mathbf{x})}\exp\left(\left(s(\mathbf{x},\mathbf{y}) - s^*(\mathbf{x},\mathbf{y})\right)/\tau\right)\right] - \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p}\left[s(\mathbf{x},\mathbf{y}) - s^*(\mathbf{x},\mathbf{y})\right].
\end{aligned}$$

This proves the desired result. □

Lemma C.3. *Let Assumption 5.1 hold. Then for any $\epsilon \in (0,\infty)$ there exists a compactly supported continuous function $u : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ such that*

$$\mathcal{L}(u) - \mathcal{L}^* \le \epsilon/2.$$

Proof. For simplicity, we denote $p_{\mathbf{x}}^+ := p(\,\cdot \mid \mathbf{x})$. Fix an arbitrary $\epsilon \in (0,\infty)$. Let Q be the probability measure on $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ defined as

$$Q(A) := \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p}\left[\mathbb{I}\left\{(\mathbf{x},\mathbf{y}) \in A\right\}\right],\ \forall \text{ measurable } A \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}.$$

In other words, Q is the zero-extension of the joint distribution of the data $(\mathbf{x},\mathbf{y}) \sim p$ to $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$. Denote $s^*(\mathbf{x},\mathbf{y}) := \tau \log \frac{p_{\mathbf{x}}^+(\mathbf{y})}{p_{\mathcal{Y}}(\mathbf{y})}$ and

$$v_0(\mathbf{x},\mathbf{y}) := \begin{cases} s^*(\mathbf{x},\mathbf{y}), & \text{if } (\mathbf{x},\mathbf{y}) \in \mathcal{X}\times\mathcal{Y}, \\ 0, & \text{if } (\mathbf{x},\mathbf{y}) \in \left(\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}\right) \setminus (\mathcal{X}\times\mathcal{Y}). \end{cases}$$

By Assumption 5.1, $v_0 \in L^1(Q)$, and there exists a sufficiently large number $M \in (1,\infty)$ such that

$$\exp(-M/\tau) + \int |v_0(\mathbf{x},\mathbf{y})| \cdot \mathbb{I}\left\{|v_0(\mathbf{x},\mathbf{y})| > M\right\} dQ(\mathbf{x},\mathbf{y}) \le \frac{\epsilon}{10\tau + 10} \tag{19}$$

Let

$$T_M : \mathbb{R} \to [-M,M], t \mapsto \begin{cases} t, & \text{if } -M \le t \le M, \\ M, & \text{if } t > M, \\ -M, & \text{if } t < -M, \end{cases}$$

and $v_1 = T_M \circ v_0$. Then it follows from (19) that

$$\|v_1 - v_0\|_{L^1(Q)} \le \int |v_0(\mathbf{x},\mathbf{y})| \cdot \mathbb{I}\left\{|v_0(\mathbf{x},\mathbf{y})| > M\right\} dQ(\mathbf{x},\mathbf{y}) < \frac{\epsilon}{10\tau+10} < \frac{\epsilon}{10}.$$

Note that

$$\begin{aligned}
&\int \exp\left(\left(M - v_0(\mathbf{x},\mathbf{y})\right)/\tau\right) dQ(\mathbf{x},\mathbf{y}) = \exp(M/\tau)\mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}\sim p_{\mathbf{x}}^+}\left[\exp\left(-v_0(\mathbf{x},\mathbf{y})/\tau\right)\right] \\
&= \exp(M/\tau)\mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}\sim p_{\mathbf{x}}^+}\frac{p_{\mathcal{Y}}(\mathbf{y})}{p_{\mathbf{x}}^+(\mathbf{y})} = \exp(M/\tau)\mathbb{E}_{\mathbf{x}}\left[\int \frac{p_{\mathcal{Y}}(\mathbf{y})}{p_{\mathbf{x}}^+(\mathbf{y})} \cdot p_{\mathbf{x}}^+(\mathbf{y}) d\mathbf{y}\right] \\
&= \exp(M/\tau)\mathbb{E}_{\mathbf{x}}\left[\int p_{\mathcal{Y}}(\mathbf{y}) d\mathbf{y}\right] = \exp(M/\tau)\mathbb{E}_{\mathbf{x}}\left[1\right] = \exp(M/\tau) < \infty.
\end{aligned}$$

By the absolute continuity of the integral, there exists a number $\delta \in (0,1)$ such that for any measurable set B with $Q(B) \leq \delta$ there holds $\int_B \exp\left(\left(M - v_0(\mathbf{x},\mathbf{y})\right)/\tau\right) dQ(\mathbf{x},\mathbf{y}) < \frac{\epsilon}{11\tau}$. According to Theorem 7.8 of Folland [1999], Q is a Radon measure on $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$. Let $\mathcal{C}_{\mathtt{c}}(\mathbb{R}^{d_1} \times \mathbb{R}^{d_2})$ be the space of all compactly supported continuous functions from $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ to $\mathbb{R}$. Then it follows from Proposition 7.9 of Folland [1999] that $\mathcal{C}_{\mathtt{c}}(\mathbb{R}^{d_1} \times \mathbb{R}^{d_2})$ is dense in $L^1(Q)$. Obviously, $v_1 \in L^1(Q)$. Thus there exists a function $v_2 \in \mathcal{C}_{\mathtt{c}}(\mathbb{R}^{d_1} \times \mathbb{R}^{d_2})$ such that $\|v_2 - v_1\|_{L^1(Q)} < \frac{\epsilon\delta}{13}$. Take $v_3 := T_M \circ v_2$. Note that $\sup_{x \neq y} \left|\frac{T_M(x) - T_M(y)}{x-y}\right| = 1$ and $T_M \circ T_M = T_M$, meaning that

$$T_M \circ v_1 = T_M \circ T_M \circ v_0 = T_M \circ v_0 = v_1.$$

Therefore, we have that

$$\|v_3 - v_1\|_{L^1(Q)} = \|T_M \circ v_2 - T_M \circ v_1\|_{L^1(Q)} \leq \sup_{x \neq y} \left|\frac{T_M(x) - T_M(y)}{x-y}\right| \cdot \|v_2 - v_1\|_{L^1(Q)}$$
$$\leq \|v_2 - v_1\|_{L^1(Q)} < \frac{\epsilon\delta}{13} \leq \frac{\epsilon}{13}.$$

Take $A := \left\{(\mathbf{x},\mathbf{y}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \;\middle|\; |v_3(\mathbf{x},\mathbf{y}) - v_1(\mathbf{x},\mathbf{y})| > \frac{\epsilon}{12}\right\}$. Then it follows from Markov's inequality that

$$Q(A) \leq \frac{\|v_3 - v_1\|_{L^1(Q)}}{\epsilon/12} \leq \frac{\epsilon\delta/13}{\epsilon/12} = \delta \cdot \frac{12}{13} < \delta.$$

Thus we have

$$\int_A \exp\left(\left(M - v_0(\mathbf{x},\mathbf{y})\right)/\tau\right) dQ(\mathbf{x},\mathbf{y}) < \frac{\epsilon}{11\tau}.$$

Denote

$$J_{\mathbf{x}} := \mathbb{E}_{\mathbf{y} \sim p^+_{\mathbf{x}}}\left[\exp\left((v_3(\mathbf{x},\mathbf{y}) - v_0(\mathbf{x},\mathbf{y}))/\tau\right) \cdot \mathbb{I}\left\{(\mathbf{x},\mathbf{y}) \notin A\right\}\right],$$
$$I_{\mathbf{x}} := \mathbb{E}_{\mathbf{y} \sim p^+_{\mathbf{x}}}\left[\exp\left((v_3(\mathbf{x},\mathbf{y}) - v_0(\mathbf{x},\mathbf{y}))/\tau\right) \cdot \mathbb{I}\left\{(\mathbf{x},\mathbf{y}) \in A\right\}\right],$$
$$H_{\mathbf{x}} := \exp\left(\tfrac{\epsilon}{12\tau}\right) \mathbb{E}_{\mathbf{y} \sim p^+_{\mathbf{x}}}\left[\exp\left((M - v_0(\mathbf{x},\mathbf{y}))/\tau\right) \cdot \mathbb{I}\left\{(\mathbf{x},\mathbf{y}) \in A\right\}\right].$$

Then we have

$$I_{\mathbf{x}} \leq \mathbb{E}_{\mathbf{y} \sim p^+_{\mathbf{x}}}\left[\exp\left((M - v_0(\mathbf{x},\mathbf{y}))/\tau\right) \cdot \mathbb{I}\left\{(\mathbf{x},\mathbf{y}) \in A\right\}\right]$$
$$\leq \exp\left(\tfrac{\epsilon}{12\tau}\right) \mathbb{E}_{\mathbf{y} \sim p^+_{\mathbf{x}}}\left[\exp\left((M - v_0(\mathbf{x},\mathbf{y}))/\tau\right) \cdot \mathbb{I}\left\{(\mathbf{x},\mathbf{y}) \in A\right\}\right] = H_{\mathbf{x}},$$

and

$$\exp\left(-\tfrac{\epsilon}{12\tau}\right) J_{\mathbf{x}} = \exp\left(-\tfrac{\epsilon}{12\tau}\right) \mathbb{E}_{\mathbf{y} \sim p^+_{\mathbf{x}}}\left[\exp\left((v_3(\mathbf{x},\mathbf{y}) - v_0(\mathbf{x},\mathbf{y}))/\tau\right) \cdot \mathbb{I}\left\{(\mathbf{x},\mathbf{y}) \notin A\right\}\right]$$
$$\leq \exp\left(-\tfrac{\epsilon}{12\tau}\right) \mathbb{E}_{\mathbf{y} \sim p^+_{\mathbf{x}}}\left[\exp\left((v_1(\mathbf{x},\mathbf{y}) + \tfrac{\epsilon}{12} - v_0(\mathbf{x},\mathbf{y}))/\tau\right) \cdot \mathbb{I}\left\{(\mathbf{x},\mathbf{y}) \notin A\right\}\right]$$
$$= \mathbb{E}_{\mathbf{y} \sim p^+_{\mathbf{x}}}\left[\exp\left((T_M(v_0(\mathbf{x},\mathbf{y})) - v_0(\mathbf{x},\mathbf{y}))/\tau\right) \cdot \mathbb{I}\left\{(\mathbf{x},\mathbf{y}) \notin A\right\}\right]$$
$$\leq \mathbb{E}_{\mathbf{y} \sim p^+_{\mathbf{x}}}\left[\exp\left((T_M(v_0(\mathbf{x},\mathbf{y})) - v_0(\mathbf{x},\mathbf{y}))/\tau\right)\right]$$
$$= \mathbb{E}_{\mathbf{y} \sim p^+_{\mathbf{x}}}\left[\exp\left((T_M(v_0(\mathbf{x},\mathbf{y})) - v_0(\mathbf{x},\mathbf{y}))/\tau\right) \cdot \mathbb{I}\left\{v_0(\mathbf{x},\mathbf{y}) \geq -M\right\}\right]$$
$$+ \mathbb{E}_{\mathbf{y} \sim p^+_{\mathbf{x}}}\left[\exp\left((T_M(v_0(\mathbf{x},\mathbf{y})) - v_0(\mathbf{x},\mathbf{y}))/\tau\right) \cdot \mathbb{I}\left\{v_0(\mathbf{x},\mathbf{y}) < -M\right\}\right]$$
$$\leq \mathbb{E}_{\mathbf{y} \sim p^+_{\mathbf{x}}}\left[\exp\left((v_0(\mathbf{x},\mathbf{y}) - v_0(\mathbf{x},\mathbf{y}))/\tau\right) \cdot \mathbb{I}\left\{v_0(\mathbf{x},\mathbf{y}) \geq -M\right\}\right]$$
$$+ \mathbb{E}_{\mathbf{y} \sim p^+_{\mathbf{x}}}\left[\exp\left((-M - v_0(\mathbf{x},\mathbf{y}))/\tau\right) \cdot \mathbb{I}\left\{v_0(\mathbf{x},\mathbf{y}) < -M\right\}\right]$$
$$\leq \mathbb{E}_{\mathbf{y} \sim p^+_{\mathbf{x}}}\left[\exp\left((v_0(\mathbf{x},\mathbf{y}) - v_0(\mathbf{x},\mathbf{y}))/\tau\right)\right] + \mathbb{E}_{\mathbf{y} \sim p^+_{\mathbf{x}}}\left[\exp\left((-M - v_0(\mathbf{x},\mathbf{y}))/\tau\right)\right]$$
$$= \mathbb{E}_{\mathbf{y} \sim p^+_{\mathbf{x}}}\left[\exp\left(0\right)\right] + \exp(-M/\tau)\mathbb{E}_{\mathbf{y} \sim p^+_{\mathbf{x}}}\left[\frac{p_{\mathcal{Y}}(\mathbf{y})}{p^+_{\mathbf{x}}(\mathbf{y})}\right]$$
$$= 1 + \exp(-M/\tau)\int \frac{p_{\mathcal{Y}}(\mathbf{y})}{p^+_{\mathbf{x}}(\mathbf{y})} \cdot p^+_{\mathbf{x}}(\mathbf{y}) d\mathbf{y} = 1 + \exp(-M/\tau)\int p_{\mathcal{Y}}(\mathbf{y}) d\mathbf{y}$$
$$= 1 + \exp(-M/\tau) \leq 1 + \tfrac{\epsilon}{10\tau + 10}$$

for any $x \in \mathcal{X}$. Take $u = v_3$. We can then use Lemma C.2 to deduce that

$$\begin{aligned}
&\mathcal{L}(u) - \mathcal{L}^* = \mathcal{L}(v_3) - \mathcal{L}^* = \mathcal{L}\big(v_3\big|_{\mathcal{X}\times\mathcal{Y}}\big) - \mathcal{L}^* \\
&= \tau\mathbb{E}_{\mathbf{x}}\left[\log\left(\mathbb{E}_{\mathbf{y}\sim p_{\mathbf{x}}^+}\exp\left((v_3(\mathbf{x},\mathbf{y}) - s^*(\mathbf{x},\mathbf{y}))/\tau\right)\right)\right] - \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p}\left[v_3(\mathbf{x},\mathbf{y}) - s^*(\mathbf{x},\mathbf{y})\right] \\
&= \tau\mathbb{E}_{\mathbf{x}}\left[\log\left(\mathbb{E}_{\mathbf{y}\sim p_{\mathbf{x}}^+}\exp\left((v_3(\mathbf{x},\mathbf{y}) - v_0(\mathbf{x},\mathbf{y}))/\tau\right)\right)\right] - \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p}\left[v_3(\mathbf{x},\mathbf{y}) - v_0(\mathbf{x},\mathbf{y})\right] \\
&= \tau\mathbb{E}_{\mathbf{x}}\left[\log\left(I_{\mathbf{x}} + J_{\mathbf{x}}\right)\right] - \int\left(v_3(\mathbf{x},\mathbf{y}) - v_0(\mathbf{x},\mathbf{y})\right)dQ(\mathbf{x},\mathbf{y}) \\
&\le \tau\mathbb{E}_{\mathbf{x}}\left[\log\left(H_{\mathbf{x}} + \exp\left(\tfrac{\epsilon}{12\tau}\right)\cdot\left(1 + \tfrac{\epsilon}{10\tau+10}\right)\right)\right] + \|v_3 - v_0\|_{L^1(Q)} \\
&= \tau\mathbb{E}_{\mathbf{x}}\left[\log\left(\mathbb{E}_{\mathbf{y}\sim p_{\mathbf{x}}^+}\left[\exp\left((M - v_0(\mathbf{x},\mathbf{y}))/\tau\right)\cdot\mathbb{I}\left\{(\mathbf{x},\mathbf{y})\in A\right\}\right] + 1 + \tfrac{\epsilon}{10\tau+10}\right)\right] \\
&\qquad + \tfrac{\epsilon}{12} + \|v_3 - v_0\|_{L^1(Q)} \\
&\le \tau\mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbf{y}\sim p_{\mathbf{x}}^+}\left[\exp\left((M - v_0(\mathbf{x},\mathbf{y}))/\tau\right)\cdot\mathbb{I}\left\{(\mathbf{x},\mathbf{y})\in A\right\}\right] + \tfrac{\epsilon}{10\tau+10}\right] \\
&\qquad + \tfrac{\epsilon}{12} + \|v_3 - v_0\|_{L^1(Q)} \\
&= \tau\int_A \exp\left((M - v_0(\mathbf{x},\mathbf{y}))/\tau\right)dQ(\mathbf{x},\mathbf{y}) + \tfrac{\epsilon\tau}{10\tau+10} + \tfrac{\epsilon}{12} + \|v_3 - v_0\|_{L^1(Q)} \\
&\le \tau\cdot\tfrac{\epsilon}{11\tau} + \tfrac{\epsilon\tau}{10\tau+10} + \tfrac{\epsilon}{12} + \|v_3 - v_1\|_{L^1(Q)} + \|v_0 - v_1\|_{L^1(Q)} \\
&\le \tfrac{\epsilon}{11} + \tfrac{\epsilon}{10} + \tfrac{\epsilon}{12} + \tfrac{\epsilon}{13} + \tfrac{\epsilon}{10} \le \epsilon/2.
\end{aligned}$$

This proves the desired result. □

Proof of Theorem 5.2. For simplicity, we denote $p_{\mathbf{x}}^+ := p(\,\cdot\,|\mathbf{x})$. Fix arbitrary positive numbers $\epsilon \in (0,\infty)$, $L_1 \in \mathbb{N}\cap(1,\infty)$, and $L_2 \in \mathbb{N}\cap(1,\infty)$. According to Lemma C.3, there exists a compactly supported continuous function $u : \mathbb{R}^{d_1}\times\mathbb{R}^{d_2} \to \mathbb{R}$, such that $\mathcal{L}(u) - \mathcal{L}^* \le \epsilon/2$. Then there exists $B \in \mathbb{N}$ such that the support of u is contained in $[-B,B]^{d_1+d_2}$ and $\sup\left\{|u(\mathbf{x},\mathbf{y})| \;\middle|\; (\mathbf{x},\mathbf{y})\in\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}\right\} < B$. For $i \in \{1,2\}$, let A_i denote the space of all continuous functions from $[-3B,3B]^{d_i}$ to $\mathbb{R}$ and $\|\cdot\|_{A_i}$ denote the uniform norm on A_i, that is, $\|w\|_{A_i} = \sup_{\mathbf{x}\in[-3B,3B]^{d_i}}|w(\mathbf{x})|$. Take

$$\mathcal{A} := \left\{\sum_{j=1}^{k} h_j(\mathbf{x})v_j(\mathbf{y}) \;\middle|\; k\in\mathbb{N},\text{ and } h_j\in A_1, v_j\in A_2,\ \forall\, j\in\{1,2,\ldots,k\}\right\}.$$

Obviously, $\mathcal{A}$ is an algebra of real continuous functions on the compact set $[-3B,3B]^{d_1+d_2}$. Moreover, $\mathcal{A}$ separates points on $[-3B,3B]^{d_1+d_2}$ and vanishes at no point of $[-3B,3B]^{d_1+d_2}$. Then it follows from the Stone–Weierstrass theorem (cf. Theorem 7.32 of Rudin [1976]) that for any continuous function w from $[-3B,3B]^{d_1+d_2}$ to $\mathbb{R}$ there exists a sequence in $\mathcal{A}$ that converges uniformly to w on $[-3B,3B]^{d_1+d_2}$. In particular, for any positive integer q, there exists

$$H_q(\mathbf{x},\mathbf{y}) = \sum_{j=1}^{k_q} h_{j,q}(\mathbf{x})v_{j,q}(\mathbf{y})$$

with $k_q\in\mathbb{N}$, $h_{j,q}\in A_1$, $v_{j,q}\in A_2$ $(j=1,2,\ldots,k_q)$ such that

$$\sup\left\{|H_q(\mathbf{x},\mathbf{y}) - u(\mathbf{x},\mathbf{y})| \;\middle|\; \mathbf{x}\in[-3B,3B]^{d_1}, \mathbf{y}\in[-3B,3B]^{d_2}\right\} \le 1/q. \tag{20}$$

It is well-known that ReLU Neural networks with one hidden layer have the universal approximation property on compact sets (see, e.g., Theorem 1 of Leshno et al. [1993]). Therefore, for any $q\in\mathbb{N}$ there exists a sufficiently large positive integer a_q and neural networks

$$\widetilde{h}_{j,q} \in \mathcal{F}_{\texttt{NN}}(d_1,1,a_q,1),\ \ \widetilde{v}_{j,q}\in\mathcal{F}_{\texttt{NN}}(d_2,1,a_q,1) \tag{21}$$

such that

$$\left\|\widetilde{h}_{j,q} - h_{j,q}\right\|_{A_1} \le \frac{1}{2qk_q}\cdot\frac{1}{1+\|v_{j,q}\|_{A_2}},\ \ \|\widetilde{v}_{j,q} - v_{j,q}\|_{A_2} \le \frac{1}{2qk_q}\cdot\frac{1}{1+\|h_{j,q}\|_{A_1}},$$

for all $j \in \{1, 2, \ldots, k_q\}$. Take

$$\widetilde{H}_q(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{k_q} \widetilde{h}_{j,q}(\mathbf{x}) \widetilde{v}_{j,q}(\mathbf{y}). \tag{22}$$

Then we have that

$$\begin{aligned}
&\left| H_q(\mathbf{x}, \mathbf{y}) - \widetilde{H}_q(\mathbf{x}, \mathbf{y}) \right| = \left| \sum_{j=1}^{k_q} \left((h_{j,q}(\mathbf{x}) - \widetilde{h}_{j,q}(\mathbf{x})) v_{j,q}(\mathbf{y}) + \widetilde{h}_{j,q}(\mathbf{x})(v_{j,q}(\mathbf{y}) - \widetilde{v}_{j,q}(\mathbf{y})) \right) \right| \\
&\leq \sum_{j=1}^{k_q} \left(\left| h_{j,q}(\mathbf{x}) - \widetilde{h}_{j,q}(\mathbf{x}) \right| |v_{j,q}(\mathbf{y})| + \left| \widetilde{h}_{j,q}(\mathbf{x}) \right| |v_{j,q}(\mathbf{y}) - \widetilde{v}_{j,q}(\mathbf{y})| \right) \\
&\leq \sum_{j=1}^{k_q} \left(\tfrac{1}{2qk_q} \cdot \frac{\|v_{j,q}\|_{A_2}}{1 + \|v_{j,q}\|_{A_2}} + \tfrac{1}{2qk_q} \cdot \frac{\left\|\widetilde{h}_{j,q}\right\|_{A_1}}{1 + \|h_{j,q}\|_{A_1}} \right) \\
&\leq \frac{1}{2qk_q} \cdot \sum_{j=1}^{k_q} \left(1 + \frac{\|h_{j,q}\|_{A_1} + \left\|h_{j,q} - \widetilde{h}_{j,q}\right\|_{A_1}}{1 + \|h_{j,q}\|_{A_1}} \right) \leq \frac{1}{2qk_q} \cdot \sum_{j=1}^{k_q} \left(1 + \frac{\|h_{j,q}\|_{A_1} + 1}{1 + \|h_{j,q}\|_{A_1}} \right) \\
&= \frac{1}{q}, \ \forall\, \mathbf{x} \in [-3B, 3B]^{d_1}, \ \mathbf{y} \in [-3B, 3B]^{d_2},
\end{aligned}$$

which, together with (20), yields

$$\left| u(\mathbf{x}, \mathbf{y}) - \widetilde{H}_q(\mathbf{x}, \mathbf{y}) \right| \leq \frac{2}{q}, \ \forall\, \mathbf{x} \in [-3B, 3B]^{d_1}, \ \mathbf{y} \in [-3B, 3B]^{d_2}. \tag{23}$$

For $i = 1, 2$ we denote

$$\begin{aligned}
&T_i : \mathbb{R}^{d_i} \to [-2B, 2B]^{d_i}, \\
&\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{d_i} \end{pmatrix} \mapsto \begin{pmatrix} \sigma(x_1 + 2B) - \sigma(x_1 - 2B) - \sigma(0x_1 + 2B) \\ \sigma(x_2 + 2B) - \sigma(x_2 - 2B) - \sigma(0x_2 + 2B) \\ \vdots \\ \sigma(x_{d_i} + 2B) - \sigma(x_{d_i} - 2B) - \sigma(0x_{d_i} + 2B) \end{pmatrix}.
\end{aligned}$$

It is easy to verify that for any $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, if $(\mathbf{x}, \mathbf{y}) \in [-2B, 2B]^{d_1+d_2}$, then $(T_1(\mathbf{x}), T_2(\mathbf{y})) = (\mathbf{x}, \mathbf{y})$, meaning that $u(\mathbf{x}, \mathbf{y}) = u(T_1(\mathbf{x}), T_2(\mathbf{y}))$; if, otherwise $(\mathbf{x}, \mathbf{y}) \notin [-2B, 2B]^{d_1+d_2}$, then $(T_1(\mathbf{x}), T_2(\mathbf{y})) \notin [-B, B]^{d_1+d_2}$, meaning that $u(T_1(\mathbf{x}), T_2(\mathbf{y})) = 0 = u(\mathbf{x}, \mathbf{y})$. Therefore, we have

$$u(T_1(\mathbf{x}), T_2(\mathbf{y})) = u(\mathbf{x}, \mathbf{y}), \ \forall\, (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}. \tag{24}$$

Define

$$\widetilde{S}_q : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}, \ (\mathbf{x}, \mathbf{y}) \mapsto \sum_{j=1}^{k_q} \widetilde{h}_{j,q}(T_1(\mathbf{x})) \widetilde{v}_{j,q}(T_2(\mathbf{y})). \tag{25}$$

Obviously,

$$T_i \in \mathcal{F}_{\mathbb{NN}}(d_i, 1, 3d_i, d_i), \ \forall\, i \in \{1, 2\}. \tag{26}$$

It follows from (23) that

$$\begin{aligned}
&\left| u(\mathbf{x}, \mathbf{y}) - \widetilde{S}_q(\mathbf{x}, \mathbf{y}) \right| \overset{\because (22), (25)}{=\!=\!=} \left| u(\mathbf{x}, \mathbf{y}) - \widetilde{H}_q(T_1(\mathbf{x}), T_2(\mathbf{y})) \right| \\
&\overset{\because (24)}{=\!=} \left| u(T_1(\mathbf{x}), T_2(\mathbf{y})) - \widetilde{H}_q(T_1(\mathbf{x}), T_2(\mathbf{y})) \right| \leq \frac{2}{q}, \ \forall\, (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}, \ \forall\, q \in \mathbb{N}.
\end{aligned}$$

Thus

$$\sup_{(\mathbf{x},\mathbf{y})\in\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}}\left|u(\mathbf{x},\mathbf{y})-\widetilde{S}_q(\mathbf{x},\mathbf{y})\right|\leq\frac{2}{q},\ \forall\ q\in\mathbb{N},\tag{27}$$

and

$$\sup_{(\mathbf{x},\mathbf{y})\in\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}}\left|\widetilde{S}_q(\mathbf{x},\mathbf{y})\right|\leq\frac{2}{q}+\sup_{(\mathbf{x},\mathbf{y})\in\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}}|u(\mathbf{x},\mathbf{y})|\leq 2+B,\ \forall\ q\in\mathbb{N}.\tag{28}$$

Combining (21), (25), (26), and (28), we get

$$\widetilde{S}_q\in\mathcal{F}_{\mathtt{S}}(2,(a_q+3d_1)\cdot k_q,2,(a_q+3d_2)\cdot k_q,k_q),\ \forall\ q\in\mathbb{N}.\tag{29}$$

From (27) we see $\left\{\widetilde{S}_q\right\}_{q=1}^{\infty}$ converges uniformly to u. We can then use Arzelà's Bounded Convergence Theorem to deduce that

$$\begin{aligned}
\mathcal{L}(u)&=\mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}\sim p_{\mathbf{x}}^{+}}\left[\tau\log\mathbb{E}_{\mathbf{y}'\sim p_{\mathcal{Y}}}\exp\left(\tfrac{u(\mathbf{x},\mathbf{y}')-u(\mathbf{x},\mathbf{y})}{\tau}\right)\right]\\
&=\mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}\sim p_{\mathbf{x}}^{+}}\left[\tau\log\lim_{q\to\infty}\mathbb{E}_{\mathbf{y}'\sim p_{\mathcal{Y}}}\exp\left(\tfrac{\widetilde{S}_q(\mathbf{x},\mathbf{y}')-\widetilde{S}_q(\mathbf{x},\mathbf{y})}{\tau}\right)\right]\\
&=\mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}\sim p_{\mathbf{x}}^{+}}\left[\lim_{q\to\infty}\tau\log\mathbb{E}_{\mathbf{y}'\sim p_{\mathcal{Y}}}\exp\left(\tfrac{\widetilde{S}_q(\mathbf{x},\mathbf{y}')-\widetilde{S}_q(\mathbf{x},\mathbf{y})}{\tau}\right)\right]\\
&=\lim_{q\to\infty}\mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}\sim p_{\mathbf{x}}^{+}}\left[\tau\log\mathbb{E}_{\mathbf{y}'\sim p_{\mathcal{Y}}}\exp\left(\tfrac{\widetilde{S}_q(\mathbf{x},\mathbf{y}')-\widetilde{S}_q(\mathbf{x},\mathbf{y})}{\tau}\right)\right]=\lim_{q\to\infty}\mathcal{L}(\widetilde{S}_q).
\end{aligned}$$

Hence there exists a sufficiently large $b\in\mathbb{N}$ such that $\left|\mathcal{L}(\widetilde{S}_b)-\mathcal{L}(u)\right|\leq\epsilon/3$. Thus

$$\mathcal{L}(\widetilde{S}_b)-\mathcal{L}^*\leq\mathcal{L}(u)-\mathcal{L}^*+\left|\mathcal{L}(\widetilde{S}_b)-\mathcal{L}(u)\right|\leq\epsilon/2+\epsilon/3<\epsilon,$$

and it follows from (29) that

$$\begin{aligned}
\widetilde{S}_b&\in\mathcal{F}_{\mathtt{S}}(2,(a_b+3d_1)\cdot k_b,2,(a_b+3d_2)\cdot k_b,k_b)\\
&\subset\mathcal{F}_{\mathtt{S}}(L_1,(a_b+3d_1)\cdot k_b,L_2,(a_b+3d_2)\cdot k_b,k_b).
\end{aligned}$$

Therefore, taking $N_1=(a_b+3d_1)\cdot k_b$, $N_2=(a_b+3d_2)\cdot k_b$, $D=k_b$, and $\tilde{s}=\widetilde{S}_b$ yields the desired result. □

D Proof of Theorem 5.4

We first present a result regarding the uniform approximation error bound for Hölder continuous functions using ReLU neural networks, which is a special case of Theorem 1.1 of Shen et al. [2020].

Lemma D.1. *Let $d\in\mathbb{N}$, $R\in(0,\infty)$, $\beta\in(0,1]$, and $B\in(0,\infty)$. Then there exist constants $C_1=C_1(d)\in\mathbb{N}$ and $C_2=C_2(d,R,\beta,B)\in(0,\infty)$ such that for any $L\in\mathbb{N}$, $N\in\mathbb{N}$ and any $f:[-R,R]^d\to\mathbb{R}$ with $\|f\|_{\beta,[-R,R]^d}\leq B$ there exists a neural network $\tilde{f}\in\mathcal{F}_{\mathtt{NN}}(d,C_1L,C_1N,1)$ such that*

$$\sup_{\mathbf{y}\in[-R,R]^d}\left|f(\mathbf{y})-\tilde{f}(\mathbf{y})\right|\leq C_2\cdot\left(\tfrac{1}{L^2N^2}\right)^{\beta/d}.\tag{30}$$

Proof. Let L, N be arbitrary positive integers, and let $f:[-R,R]^d\to\mathbb{R}$ be an arbitrary function satisfying $\|f\|_{\beta,[-R,R]^d}\leq B$. Denote $\mathbf{1}:=(1,1,\ldots,1)\in\mathbb{R}^d$, and define

$$g:[0,1]^d\to\mathbb{R},\ \mathbf{x}\mapsto f(2R\mathbf{x}-R\mathbf{1}).$$

Obviously, the modulus of continuity of g satisfies

$$\begin{aligned}
\omega_g(r)&:=\sup\left\{|g(\mathbf{x})-g(\mathbf{y})|\ \middle|\ \mathbf{x},\mathbf{y}\in[0,1]^d,\|\mathbf{x}-\mathbf{y}\|_2\leq r\right\}\\
&=\sup\left\{|f(2R\mathbf{x}-R\mathbf{1})-f(2R\mathbf{y}-R\mathbf{1})|\ \middle|\ \mathbf{x},\mathbf{y}\in[0,1]^d,\|\mathbf{x}-\mathbf{y}\|_2\leq r\right\}\\
&\leq\sup\left\{B\cdot\|(2R\mathbf{x}-R\mathbf{1})-(2R\mathbf{y}-R\mathbf{1})\|_\infty^\beta\ \middle|\ \mathbf{x},\mathbf{y}\in[0,1]^d,\|\mathbf{x}-\mathbf{y}\|_2\leq r\right\}\\
&=\sup\left\{B\cdot(2R)^\beta\|\mathbf{x}-\mathbf{y}\|_\infty^\beta\ \middle|\ \mathbf{x},\mathbf{y}\in[0,1]^d,\|\mathbf{x}-\mathbf{y}\|_2\leq r\right\}\\
&\leq B\cdot(2R)^\beta\cdot r^\beta,\ \forall\ r\in(0,\infty).
\end{aligned}$$

Then it follows from Theorem 1.1 of Shen et al. [2020] that there exists

$$\tilde{g} \in \mathcal{F}_{\texttt{NN}}(d, 12L + 14 + 2d, 3^{d+3} \max\{dN, N+1\} + 1, 1)$$

such that

$$\sup_{\mathbf{x}\in[0,1]^d} |g(\mathbf{x}) - \tilde{g}(\mathbf{x})| \leq 19\sqrt{d} \cdot B \cdot (2R)^\beta \cdot \left(N^{-2/d} L^{-2/d}\right)^\beta.$$

Take $C_1 := 3^{d+3} \cdot (1+d) + 1$, $C_2 := 19\sqrt{d} \cdot B \cdot (2R)^\beta$, and $\tilde{f}(\mathbf{y}) := \tilde{g}(\mathbf{y}/(2R) + \frac{1}{2}\mathbf{1})$. Then we have

$$\begin{aligned}
\tilde{f} &\in \mathcal{F}_{\texttt{NN}}(d, 12L + 14 + 2d, 3^{d+3} \max\{dN, N+1\} + 1, 1) \\
&\subset \mathcal{F}_{\texttt{NN}}(d, (26+2d) \cdot L, 3^{d+3} \cdot (d+1) \cdot N + 1, 1) \subset \mathcal{F}_{\texttt{NN}}(d, C_1 L, C_1 N, 1),
\end{aligned}$$

and

$$\begin{aligned}
&\sup_{\mathbf{y}\in[-R,R]^d} \left| f(\mathbf{y}) - \tilde{f}(\mathbf{y}) \right| = \sup_{\mathbf{x}\in[0,1]^d} \left| f(2R\mathbf{x} - R\mathbf{1}) - \tilde{f}(2R\mathbf{x} - R\mathbf{1}) \right| \\
&= \sup_{\mathbf{x}\in[0,1]^d} |g(\mathbf{x}) - \tilde{g}(\mathbf{x})| \leq 19\sqrt{d} \cdot B \cdot (2R)^\beta \cdot \left(N^{-2/d} L^{-2/d}\right)^\beta = C_2 \cdot \left(\tfrac{1}{L^2 N^2}\right)^{\beta/d}.
\end{aligned}$$

This completes the proof. □

Lemma D.2. *Let Assumption 5.1 hold. Denote $s^*(\mathbf{x}, \mathbf{y}) := \tau \log \frac{p(\mathbf{y}|\mathbf{x})}{p_{\mathcal{Y}}(\mathbf{y})}$. Then for any bounded measurable function $u : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ there holds*

$$\mathcal{L}(u) - \mathcal{L}^* \leq \mathbb{E}_{\mathbf{x}} \left[\sup_{\mathbf{y}\in\mathcal{Y}} \tfrac{|u(\mathbf{x},\mathbf{y}) - s^*(\mathbf{x},\mathbf{y})|^2}{2\tau} \right].$$

Proof. For simplicity, we denote $p^+_{\mathbf{x}} := p(\,\cdot \mid \mathbf{x})$. Let u be an arbitrary bounded measurable function from $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$. Define

$$a(\mathbf{x}) := \sup_{\mathbf{y}\in\mathcal{Y}} |u(\mathbf{x}, \mathbf{y}) - s^*(\mathbf{x}, \mathbf{y})|, \ \forall\, \mathbf{x} \in \mathcal{X}, \tag{31}$$

and

$$H(\mathbf{x}, \mathbf{y}) := u(\mathbf{x}, \mathbf{y}) - s^*(\mathbf{x}, \mathbf{y}), \ \forall\, \mathbf{x} \in \mathcal{X}, \ \mathbf{y} \in \mathcal{Y}.$$

Then define

$$v(\mathbf{x}) := \begin{cases} \mathbb{E}_{\mathbf{y}\sim p^+_{\mathbf{x}}} H(\mathbf{x}, \mathbf{y}), & \text{if } \mathbb{E}_{\mathbf{y}\sim p^+_{\mathbf{x}}} |H(\mathbf{x}, \mathbf{y})| < \infty, \\ 0, & \text{if } \mathbb{E}_{\mathbf{y}\sim p^+_{\mathbf{x}}} |H(\mathbf{x}, \mathbf{y})| = \infty \end{cases}$$

for all $\mathbf{x} \in \mathcal{X}$, and $A := \{\mathbf{x} \in \mathcal{X} \mid a(\mathbf{x}) < \infty\}$. It follows from Assumption 5.1 and the boundedness of u that

$$\mathbb{E}_{\mathbf{x}} |v(\mathbf{x})| \leq \mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{y}\sim p^+_{\mathbf{x}}} |H(\mathbf{x}, \mathbf{y})| < \infty,$$

which means that $\mathbb{E}_{\mathbf{y}\sim p^+_{\mathbf{x}}} |H(\mathbf{x}, \mathbf{y})| < \infty$ for almost all $\mathbf{x} \in \mathcal{X}$, and

$$\mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{y}\sim p^+_{\mathbf{x}}} H(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{\mathbf{x}} v(\mathbf{x}). \tag{32}$$

According to Hoeffding's lemma (see, e.g., Lemma 2.2 Boucheron et al. [2013]),

$$\begin{aligned}
&- v(\mathbf{x}) \cdot \lambda + \log \mathbb{E}_{\mathbf{y}\sim p^+_{\mathbf{x}}} \exp\left(H(\mathbf{x}, \mathbf{y}) \cdot \lambda\right) = \log \mathbb{E}_{\mathbf{y}\sim p^+_{\mathbf{x}}} \exp\left((H(\mathbf{x}, \mathbf{y}) - v(\mathbf{x})) \cdot \lambda\right) \\
&\leq \frac{\lambda^2 \cdot (a(\mathbf{x}) - v(\mathbf{x}) - (-a(\mathbf{x}) - v(\mathbf{x})))^2}{8} = \frac{\lambda^2 \cdot a(\mathbf{x})^2}{2}, \ \forall\, \lambda \in \mathbb{R}, \ \forall\, \mathbf{x} \in A.
\end{aligned}$$

Taking $\lambda = 1/\tau$ above, we get

$$-v(\mathbf{x}) + \tau \log \mathbb{E}_{\mathbf{y}\sim p^+_{\mathbf{x}}} \exp\left(H(\mathbf{x}, \mathbf{y})/\tau\right) \leq \frac{a(\mathbf{x})^2}{2\tau}, \ \forall\, \mathbf{x} \in A. \tag{33}$$

Furthermore,

$$-v(\mathbf{x}) + \tau \log \mathbb{E}_{\mathbf{y}\sim p_{\mathbf{x}}^+} \exp\left(H(\mathbf{x},\mathbf{y})/\tau\right) \leq +\infty = \frac{a(\mathbf{x})^2}{2\tau},\ \forall\, \mathbf{x} \in \mathcal{X} \setminus A. \tag{34}$$

Combining (33) and (34), we obtain

$$-v(\mathbf{x}) + \tau \log \mathbb{E}_{\mathbf{y}\sim p_{\mathbf{x}}^+} \exp\left(H(\mathbf{x},\mathbf{y})/\tau\right) \leq \frac{a(\mathbf{x})^2}{2\tau},\ \forall\, \mathbf{x} \in \mathcal{X}.$$

Then it follows from Lemma C.2

$$\begin{aligned}
&\mathcal{L}(u) - \mathcal{L}^* = \mathbb{E}_{\mathbf{x}}\left[\tau \log \mathbb{E}_{\mathbf{y}\sim p_{\mathbf{x}}^+} \exp\left((u(\mathbf{x},\mathbf{y}) - s^*(\mathbf{x},\mathbf{y}))/\tau\right)\right] - \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}\sim p_{\mathbf{x}}^+}\left[u(\mathbf{x},\mathbf{y}) - s^*(\mathbf{x},\mathbf{y})\right] \\
&= \mathbb{E}_{\mathbf{x}}\left[\tau \log \mathbb{E}_{\mathbf{y}\sim p_{\mathbf{x}}^+} \exp\left(H(\mathbf{x},\mathbf{y})/\tau\right)\right] - \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{y}\sim p_{\mathbf{x}}^+}\left[H(\mathbf{x},\mathbf{y})\right] \\
&\overset{\because (32)}{=\!=} \mathbb{E}_{\mathbf{x}}\left[\tau \log \mathbb{E}_{\mathbf{y}\sim p_{\mathbf{x}}^+} \exp\left(H(\mathbf{x},\mathbf{y})/\tau\right)\right] - \mathbb{E}_{\mathbf{x}}\left[v(\mathbf{x})\right] \\
&= \mathbb{E}_{\mathbf{x}}\left[-v(\mathbf{x}) + \tau \log \mathbb{E}_{\mathbf{y}\sim p_{\mathbf{x}}^+} \exp\left(H(\mathbf{x},\mathbf{y})/\tau\right)\right] \\
&\leq \mathbb{E}_{\mathbf{x}}\left[\frac{a(\mathbf{x})^2}{2\tau}\right] \overset{\because (31)}{=\!=} \mathbb{E}_{\mathbf{x}}\left[\sup_{\mathbf{y}\in\mathcal{Y}} \tfrac{|u(\mathbf{x},\mathbf{y}) - s^*(\mathbf{x},\mathbf{y})|^2}{2\tau}\right].
\end{aligned}$$

This proves the desired result. □

Lemma D.3. *Let $\beta \in (0,1]$, $d \in \mathbb{N}$, $\varnothing \subsetneqq U \subset V \subset \mathbb{R}^d$, and f be a function from U to $\mathbb{R}$. Then there exists a function $g : V \to \mathbb{R}$ such that $\|g\|_{\beta,V} \leq \|f\|_{\beta,U}$ and*

$$f(\mathbf{x}) = g(\mathbf{x}),\ \forall\, \mathbf{x} \in U.$$

Proof. The case $\|f\|_{\beta,U} = \infty$ is trivial. In what follows, we assume $\|f\|_{\beta,U} < \infty$. Let

$$\rho : \mathbb{R}^d \times \mathbb{R}^d \to [0,\infty),\ (\mathbf{x},\mathbf{y}) \mapsto \|\mathbf{x}-\mathbf{y}\|_\infty^\beta,$$

and $a := \|f\|_{\beta,U}$. Then it is easy to verify that ρ is a metric on $\mathbb{R}^d$, and

$$\sup\left\{ \tfrac{|f(\mathbf{x})-f(\mathbf{y})|}{\|\mathbf{x}-\mathbf{y}\|_\infty^\beta} \;\middle|\; U \ni \mathbf{x} \neq \mathbf{y} \in U \right\} \leq \|f\|_{\beta,U} = a.$$

Therefore, f is a-Lipschitz continuous on U with respect to the metric ρ, that is,

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq a \cdot \rho(\mathbf{x},\mathbf{y}),\ \forall\, \mathbf{x} \in U,\ \mathbf{y} \in U.$$

Then it follows from Theorem 1 of McShane [1934] that f can be extended to $h : \mathbb{R}^d \to \mathbb{R}$ preserving the same Lipschitz condition, that is,

$$h(\mathbf{x}) = f(\mathbf{x}),\ \forall\, \mathbf{x} \in U, \tag{35}$$

and

$$|h(\mathbf{x}) - h(\mathbf{y})| \leq a \cdot \rho(\mathbf{x},\mathbf{y}),\ \forall\, \mathbf{x} \in \mathbb{R}^d,\ \mathbf{y} \in \mathbb{R}^d. \tag{36}$$

Let $T : \mathbb{R} \to \mathbb{R},\ t \mapsto \min\left\{\max\left\{t, -a\right\}, a\right\}$. Obviously,

$$|T(x) - T(y)| \leq |x-y|, \forall\, x \in \mathbb{R},\ \forall\, y \in \mathbb{R}, \tag{37}$$

and

$$T(x) = x,\ \forall\, x \in [-a,a]. \tag{38}$$

Note that $|f(\mathbf{x})| \leq \|f\|_{\beta,U} = a$ for all $\mathbf{x} \in U$. Therefore,

$$T(h(\mathbf{x})) \overset{\because (35)}{=\!=} T(f(\mathbf{x})) \overset{\because (38)}{=\!=} f(\mathbf{x}),\ \forall\, \mathbf{x} \in U. \tag{39}$$

Combining (36) and (37), we get

$$|T(h(\mathbf{x})) - T(h(\mathbf{y}))| \leq |h(\mathbf{x}) - h(\mathbf{y})| \leq a \cdot \rho(\mathbf{x},\mathbf{y}) = a \cdot \|\mathbf{x}-\mathbf{y}\|_\infty^\beta,\ \forall\, \mathbf{x},\mathbf{y} \in \mathbb{R}^d.$$

Thus

$$\sup_{V\ni\mathbf{x}\neq\mathbf{y}\in V}\tfrac{|T(h(\mathbf{x}))-T(h(\mathbf{y}))|}{\|\mathbf{x}-\mathbf{y}\|_\infty^\beta}\leq a.$$

Therefore,

$$\begin{aligned}\|T\circ h\|_{\beta,V}&=\max\left\{\sup_{\mathbf{x}\in V}|T(h(\mathbf{x}))|\,,\sup_{V\ni\mathbf{x}\neq\mathbf{y}\in V}\tfrac{|T(h(\mathbf{x}))-T(h(\mathbf{y}))|}{\|\mathbf{x}-\mathbf{y}\|_\infty^\beta}\right\}\\&\leq\max\left\{\sup_{\mathbf{x}\in V}|T(h(\mathbf{x}))|\,,a\right\}\leq\max\left\{\sup_{t\in\mathbb{R}}|T(t)|\,,a\right\}=\max\{a,a\}=a=\|f\|_{\beta,U}\,,\end{aligned}$$

which, together with (39), implies that the function $g:V\to\mathbb{R},\mathbf{x}\mapsto T(h(\mathbf{x}))$ has all the desired properties. This completes the proof. □

Lemma D.4. *Suppose $m_1\in\mathbb{N}$, $m_2\in\mathbb{N}$, $R\in(0,\infty)$, $\beta\in(0,1]$, $B\in(0,\infty)$. Fix arbitrary positive numbers θ_1, θ_2, θ_3, θ_4, θ_5. Then for sufficiently small $\epsilon\in(0,\infty)$ and $L_1\asymp\epsilon^{-\theta_1}$, $N_1\asymp\epsilon^{-\theta_2}$, $L_2\asymp\epsilon^{-\theta_3}$, $N_2\asymp\epsilon^{-\theta_4}$, $D\asymp\epsilon^{-\theta_5}$ there holds*

$$\begin{aligned}&\sup_{\substack{h:[-R,R]^{m_1+m_2}\to\mathbb{R}\\\|h\|_{\beta,[-R,R]^{m_1+m_2}}\leq B}}\inf\left\{\sup_{\substack{\mathbf{x}\in[-R,R]^{m_1}\\\mathbf{y}\in[-R,R]^{m_2}}}\left|f(\mathbf{x})^\top g(\mathbf{y})-h(\mathbf{x},\mathbf{y})\right|\;\middle|\;\begin{array}{c}f\in\mathcal{F}_{\mathtt{NN}}(m_1,L_1,N_1,D),\\g\in\mathcal{F}_{\mathtt{NN}}(m_2,L_2,N_2,D)\end{array}\right\}\\&\lesssim\left|\epsilon^{\theta_3+\theta_4-\min\{\theta_2,\theta_4,\theta_5\}}\right|^{2\beta/m_2}+\left|\epsilon^{\min\{\theta_2,\theta_4,\theta_5\}}\right|^{\beta/m_1}.\end{aligned}$$

Proof. Let $h:[-R,R]^{m_1+m_2}\to\mathbb{R}$ be an arbitrary function that satisfies $\|h\|_{\beta,[-R,R]^{m_1+m_2}}\leq B$. Obviously,

$$\|h(\mathbf{a},\cdot)\|_{\beta,[-R,R]^{m_2}}\leq\|h\|_{\beta,[-R,R]^{m_1+m_2}}\leq B,\ \forall\,\mathbf{a}\in[-R,R]^{m_1}.$$

Then it follows from Lemma D.1 that there exist constants $C_2=C_2(m_2,R,\beta,B)\in(0,\infty)$ and $C_1=C_1(m_2)\in\mathbb{N}$ such that for any $L\in\mathbb{N}$, $N\in\mathbb{N}$ and any $\mathbf{a}\in[-R,R]^{m_1}$ there exists a neural network

$$\tilde{f}_{\mathbf{a},L,N}\in\mathcal{F}_{\mathtt{NN}}(d,C_1L,C_1N,1)\tag{40}$$

such that

$$\sup_{\mathbf{y}\in[-R,R]^{m_2}}\left|h(\mathbf{a},\mathbf{y})-\tilde{f}_{\mathbf{a},L,N}(\mathbf{y})\right|\leq C_2\cdot\left(\tfrac{1}{L^2N^2}\right)^{\beta/m_2}.\tag{41}$$

Let

$$G:=\left\lfloor\left(\min\left\{D^{1/m_1},\left(\tfrac{N_1}{6m_1}\right)^{1/m_1},\left(\tfrac{N_2}{2C_1}\right)^{1/m_1}\right\}-1\right)\cdot\frac{1}{2}\right\rfloor\asymp\epsilon^{-\min\{\theta_2,\theta_4,\theta_5\}/m_1}\tag{42}$$

and $N_3:=\lfloor N_2/\left(C_1\cdot(2G+1)^{m_1}\right)\rfloor$. Obviously,

$$2\leq\lfloor N_2/\left(C_1\cdot(2G+1)^{m_1}\right)\rfloor=N_3\asymp\epsilon^{\min\{\theta_2,\theta_4,\theta_5\}-\theta_4}.\tag{43}$$

We then take

$$L_3:=\sup\left\{L\in\mathbb{N}\cap[2,\infty)\mid C_1L\leq L_2\right\}.\tag{44}$$

From $L_2\asymp\epsilon^{-\theta_3}$ and $N_3\asymp\epsilon^{\min\{\theta_2,\theta_4,\theta_5\}-\theta_4}$ we get $L_3\asymp\epsilon^{-\theta_3}$. For any $k\in\mathbb{Z}$, let $l_k:\mathbb{R}\to[0,1]$ be the hat function which is linear on $[\frac{R\cdot(k-1)}{G},\frac{R\cdot k}{G})]$ and $[\frac{R\cdot k}{G},\frac{R\cdot(k+1)}{G})]$ respectively, and satisfies

$$l_k(\tfrac{R\cdot k}{G})=1\geq0=l_k(x),\ \forall\,x\in\mathbb{R}\setminus(\tfrac{R\cdot(k-1)}{G},\tfrac{R\cdot(k+1)}{G}).\tag{45}$$

Obviously, l_k can be expressed as

$$l_k(x)=\left(\sigma(Gx-Rk+R)+\sigma(Gx-Rk-R)-2\sigma(Gx-Rk)\right)/R,\ \forall\,x\in\mathbb{R},\tag{46}$$

and $\{l_k\}_{k=-G}^{G}$ forms a partition of unity on $[R,R]$, that is,

$$\sum_{k=-G}^{G} l_k(x) = 1,\ \forall\, x \in [-R,R]. \tag{47}$$

Let

$$\Omega := \left\{ \tfrac{Rk}{G} \;\middle|\; k \in [-G,G] \cap \mathbb{Z} \right\} \subset [-R,R].$$

Then for any $\mathbf{a} = (\frac{Rk_1}{G}, \frac{Rk_2}{G}, \ldots, \frac{Rk_{m_1}}{G}) \in \Omega^{m_1}$, define

$$u_{\mathbf{a}} : \mathbb{R}^{m_1} \to [0,1],\ \mathbf{x} = (x_1, \ldots, x_{m_1}) \mapsto \prod_{j=1}^{m_1} l_{k_j}(x_j).$$

Then we have that

$$u_{\mathbf{a}}(\mathbf{x}) \overset{\because (45)}{=\!=} 0,\ \forall\, \mathbf{a} \in \Omega^{m_1}, \forall\, \mathbf{x} \in \left\{ \mathbf{z} \in \mathbb{R}^{m_1} \mid \|\mathbf{z} - \mathbf{a}\|_\infty \geq R/G \right\}, \tag{48}$$

and

$$\begin{aligned}
&\sum_{\mathbf{a} \in \Omega^{m_1}} u_{\mathbf{a}}(\mathbf{x}) = \sum_{k_1=-G}^{G} \sum_{k_2=-G}^{G} \cdots \sum_{k_{m_1}=-G}^{G} \prod_{j=1}^{m_1} l_{k_j}(x_j) \\
&= \prod_{j=1}^{m_1} \sum_{k_j=-G}^{G} l_{k_j}(x_j) \overset{\because (47)}{=\!=} \prod_{j=1}^{m_1} 1 = 1,\ \forall\, \mathbf{x} = (x_1, \ldots, x_{m_1}) \in [-R,R]^{m_1}.
\end{aligned}$$

Thus

$$\sum_{\mathbf{a} \in \Omega^{m_1}} u_{\mathbf{a}}(\mathbf{x}) = 1,\ \forall\, \mathbf{x} \in [-R,R]^{m_1}. \tag{49}$$

Let $L_4 := \left\lfloor \frac{L_1 - 2}{\lceil \log_2 m_1 \rceil} \right\rfloor - 5 \asymp \epsilon^{-\theta_1}$. According to Lemma A.3 of Schmidt-Hieber [2020], there exists a neural network

$$H \in \mathcal{F}_{\mathtt{NN}}(m_1, (L_4 + 5) \cdot \lceil \log_2 m_1 \rceil, 6m_1, 1) \tag{50}$$

such that

$$\left| H(x_1, \ldots, x_{m_1}) - \prod_{j=1}^{m_1} x_j \right| \leq m_1^2 \cdot 2^{-L_4},\ \forall\, x_1, x_2, \ldots, x_{m_1} \in [0,1].$$

For each $\mathbf{a} = (\frac{Rk_1}{G}, \frac{Rk_2}{G}, \ldots, \frac{Rk_{m_1}}{G}) \in \Omega^{m_1}$, define

$$\tilde{u}_{\mathbf{a}} : \mathbb{R}^{m_1} \to \mathbb{R}, \mathbf{x} = (x_1, \ldots, x_{m_1}) \mapsto H(l_{k_1}(x_1), l_{k_2}(x_2), \ldots, l_{k_{m_1}}(x_{m_1})).$$

Then we have that

$$\begin{aligned}
&|\tilde{u}_{\mathbf{a}}(\mathbf{x}) - u_{\mathbf{a}}(\mathbf{x})| = \left| H(l_{k_1}(x_1), l_{k_2}(x_2), \ldots, l_{k_{m_1}}(x_{m_1})) - \prod_{j=1}^{m_1} l_{k_j}(x_j) \right| \\
&\leq m_1^2 \cdot 2^{-L_4},\ \forall\, \mathbf{x} = (x_1, \ldots, x_{m_1}) \in \mathbb{R}^{m_1}.
\end{aligned}$$

Thus

$$|\tilde{u}_{\mathbf{a}}(\mathbf{x}) - u_{\mathbf{a}}(\mathbf{x})| \leq m_1^2 \cdot 2^{-L_4},\ \forall\, \mathbf{x} \in \mathbb{R}^{m_1},\ \forall\, \mathbf{a} \in \Omega^{m_1}. \tag{51}$$

Combining (46) and (50), we obtain that

$$\begin{aligned}
&\tilde{u}_{\mathbf{a}} \in \mathcal{F}_{\mathtt{NN}}(m_1, 2 + (L_4 + 5) \cdot \lceil \log_2 m_1 \rceil, 6m_1, 1) \\
&= \mathcal{F}_{\mathtt{NN}}(m_1, 2 + \left\lfloor \tfrac{L_1 - 2}{\lceil \log_2 m_1 \rceil} \right\rfloor \cdot \lceil \log_2 m_1 \rceil, 6m_1, 1) \subset \mathcal{F}_{\mathtt{NN}}(m_1, L_1, 6m_1, 1),\ \forall\, \mathbf{a} \in \Omega^{m_1}.
\end{aligned}$$

Thus

$$\tilde{u}_{\mathbf{a}} \in \mathcal{F}_{\mathbb{NN}}(m_1, L_1, 6m_1, 1),\ \forall\ \mathbf{a} \in \Omega^{m_1}. \tag{52}$$

Let

$$q : \mathbb{R}^{m_1} \times [-R, R]^{m_2} \to \mathbb{R}, (\mathbf{x}, \mathbf{y}) \mapsto \sum_{\mathbf{a} \in \Omega^{m_1}} u_{\mathbf{a}}(\mathbf{x}) h(\mathbf{a}, \mathbf{y}),$$

and

$$\tilde{q} : \mathbb{R}^{m_1} \times \mathbb{R}^{m_2} \to \mathbb{R}, (\mathbf{x}, \mathbf{y}) \mapsto \sum_{\mathbf{a} \in \Omega^{m_1}} \tilde{u}_{\mathbf{a}}(\mathbf{x}) \tilde{f}_{\mathbf{a}, L_3, N_3}(\mathbf{y}).$$

Obviously,

$$\begin{aligned}
&|q(\mathbf{x}, \mathbf{y}) - h(\mathbf{x}, \mathbf{y})| \overset{\because (49)}{=\!=\!=} \left| \sum_{\mathbf{a} \in \Omega^{m_1}} u_{\mathbf{a}}(\mathbf{x}) h(\mathbf{a}, \mathbf{y}) - \sum_{\mathbf{a} \in \Omega^{m_1}} u_{\mathbf{a}}(\mathbf{x}) h(\mathbf{x}, \mathbf{y}) \right| \\
&= \left| \sum_{\mathbf{a} \in \Omega^{m_1}} u_{\mathbf{a}}(\mathbf{x}) (h(\mathbf{a}, \mathbf{y}) - h(\mathbf{x}, \mathbf{y})) \right| \leq \sum_{\mathbf{a} \in \Omega^{m_1}} u_{\mathbf{a}}(\mathbf{x}) \left| h(\mathbf{a}, \mathbf{y}) - h(\mathbf{x}, \mathbf{y}) \right| \\
&\overset{\because (48)}{=\!=\!=} \sum_{\mathbf{a} \in \Omega^{m_1}, \|\mathbf{x} - \mathbf{a}\|_\infty \leq \frac{R}{G}} u_{\mathbf{a}}(\mathbf{x}) \left| h(\mathbf{a}, \mathbf{y}) - h(\mathbf{x}, \mathbf{y}) \right| \leq \sum_{\mathbf{a} \in \Omega^{m_1}, \|\mathbf{x} - \mathbf{a}\|_\infty \leq \frac{R}{G}} u_{\mathbf{a}}(\mathbf{x}) \cdot B \cdot \|\mathbf{a} - \mathbf{x}\|_\infty^\beta \\
&\leq \sum_{\mathbf{a} \in \Omega^{m_1}, \|\mathbf{x} - \mathbf{a}\|_\infty \leq \frac{R}{G}} u_{\mathbf{a}}(\mathbf{x}) \cdot B \cdot \left(\tfrac{R}{G}\right)^\beta \leq \sum_{\mathbf{a} \in \Omega^{m_1}} u_{\mathbf{a}}(\mathbf{x}) \cdot B \cdot \left(\tfrac{R}{G}\right)^\beta \\
&\overset{\because (49)}{=\!=\!=} B \cdot \left(\tfrac{R}{G}\right)^\beta,\ \forall\ (\mathbf{x}, \mathbf{y}) \in [-R, R]^{m_1} \times [-R, R]^{m_2}.
\end{aligned}$$

Thus

$$\sup_{\substack{\mathbf{x} \in [-R, R]^{m_1} \\ \mathbf{y} \in [-R, R]^{m_2}}} |q(\mathbf{x}, \mathbf{y}) - h(\mathbf{x}, \mathbf{y})| \leq B \cdot \left(\tfrac{R}{G}\right)^\beta. \tag{53}$$

Combining (41) and (51), we obtain that

$$\begin{aligned}
&|q(\mathbf{x}, \mathbf{y}) - \tilde{q}(\mathbf{x}, \mathbf{y})| = \left| \sum_{\mathbf{a} \in \Omega^{m_1}} u_{\mathbf{a}}(\mathbf{x}) h(\mathbf{a}, \mathbf{y}) - \sum_{\mathbf{a} \in \Omega^{m_1}} \tilde{u}_{\mathbf{a}}(\mathbf{x}) \tilde{f}_{\mathbf{a}, L_3, N_3}(\mathbf{y}) \right| \\
&= \left| \sum_{\mathbf{a} \in \Omega^{m_1}} u_{\mathbf{a}}(\mathbf{x}) (h(\mathbf{a}, \mathbf{y}) - \tilde{f}_{\mathbf{a}, L_3, N_3}(\mathbf{y})) + \sum_{\mathbf{a} \in \Omega^{m_1}} (u_{\mathbf{a}}(\mathbf{x}) - \tilde{u}_{\mathbf{a}}(\mathbf{x})) \tilde{f}_{\mathbf{a}, L_3, N_3}(\mathbf{y}) \right| \\
&\leq \sum_{\mathbf{a} \in \Omega^{m_1}} u_{\mathbf{a}}(\mathbf{x}) \left| h(\mathbf{a}, \mathbf{y}) - \tilde{f}_{\mathbf{a}, L_3, N_3}(\mathbf{y}) \right| + \sum_{\mathbf{a} \in \Omega^{m_1}} |u_{\mathbf{a}}(\mathbf{x}) - \tilde{u}_{\mathbf{a}}(\mathbf{x})| \left| \tilde{f}_{\mathbf{a}, L_3, N_3}(\mathbf{y}) \right| \\
&\leq \sum_{\mathbf{a} \in \Omega^{m_1}} u_{\mathbf{a}}(\mathbf{x}) \left| C_2 \cdot \left(\tfrac{1}{L_3^2 N_3^2}\right)^{\beta/m_2} \right| + \sum_{\mathbf{a} \in \Omega^{m_1}} |u_{\mathbf{a}}(\mathbf{x}) - \tilde{u}_{\mathbf{a}}(\mathbf{x})| \left| |h(\mathbf{a}, \mathbf{y})| + C_2 \cdot \left(\tfrac{1}{L^2 N^2}\right)^{\beta/m_2} \right| \\
&\leq \sum_{\mathbf{a} \in \Omega^{m_1}} u_{\mathbf{a}}(\mathbf{x}) \left| C_2 \cdot \left(\tfrac{1}{L_3^2 N_3^2}\right)^{\beta/m_2} \right| + \sum_{\mathbf{a} \in \Omega^{m_1}} \left| m_1^2 \cdot 2^{-L_4} \right| \left| B + C_2 \cdot \left(\tfrac{1}{L^2 N^2}\right)^{\beta/m_2} \right| \\
&\overset{\because (49)}{=\!=\!=} C_2 \cdot \left(\tfrac{1}{L_3^2 N_3^2}\right)^{\beta/m_2} + (2G+1)^{m_1} \cdot \left| m_1^2 \cdot 2^{-L_4} \right| \left| B + C_2 \cdot \left(\tfrac{1}{L^2 N^2}\right)^{\beta/m_2} \right| \\
&< \infty,\ \forall\ (\mathbf{x}, \mathbf{y}) \in [-R, R]^{m_1} \times [-R, R]^{m_2},
\end{aligned}$$

which, together with (53), yields

$$\begin{aligned}
&\sup_{[-R, R]^{m_1 + m_2}} |\tilde{q} - h| \leq \sup_{[-R, R]^{m_1 + m_2}} |\tilde{q} - q| + \sup_{[-R, R]^{m_1 + m_2}} |q - h| \\
&\leq C_2 \cdot \left(\tfrac{1}{L_3^2 N_3^2}\right)^{\beta/m_2} + (2G+1)^{m_1} \cdot \left| m_1^2 \cdot 2^{-L_4} \right| \left| B + C_2 \cdot \left(\tfrac{1}{L_3^2 N_3^2}\right)^{\beta/m_2} \right| + B \cdot \left(\tfrac{R}{G}\right)^\beta < \infty.
\end{aligned}$$

From (40), (42), (43), (44), and (52) we see there exist

$$\begin{aligned}\tilde{v} &\in \mathcal{F}_{\mathtt{NN}}(m_1, L_1, 6m_1 \cdot |\Omega^{m_1}|, |\Omega^{m_1}|) = \mathcal{F}_{\mathtt{NN}}(m_1, L_1, 6m_1 \cdot (2G+1)^{m_1}, (2G+1)^{m_1}) \\ &\subset \mathcal{F}_{\mathtt{NN}}(m_1, L_1, N_1, (2G+1)^{m_1})\end{aligned}$$

and

$$\begin{aligned}\tilde{g} &\in \mathcal{F}_{\mathtt{NN}}(m_2, C_1 L_3, C_1 N_3 \cdot |\Omega^{m_1}|, |\Omega^{m_1}|) \\ &\subset \mathcal{F}_{\mathtt{NN}}(m_2, L_2, C_1 N_3 \cdot (2G+1)^{m_1}, (2G+1)^{m_1}) \subset \mathcal{F}_{\mathtt{NN}}(m_2, L_2, N_2, (2G+1)^{m_1})\end{aligned}$$

such that

$$\tilde{q}(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{a} \in \Omega^{m_1}} \tilde{u}_{\mathbf{a}}(\mathbf{x}) \tilde{f}_{\mathbf{a}, L_3, N_3}(\mathbf{y}) = \tilde{v}(\mathbf{x})^\top \tilde{g}(\mathbf{y}),\ \forall\ (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{m_1} \times \mathbb{R}^{m_2},$$

which, together with the fact that $(2G+1)^{m_1} \leq D$, yields

$$\begin{aligned}&\inf \left\{ \sup_{\substack{\mathbf{x} \in [-R,R]^{m_1} \\ \mathbf{y} \in [-R,R]^{m_2}}} \left| f(\mathbf{x})^\top g(\mathbf{y}) - h(\mathbf{x}, \mathbf{y}) \right| \,\middle|\, \begin{array}{l} f \in \mathcal{F}_{\mathtt{NN}}(m_1, L_1, N_1, D), \\ g \in \mathcal{F}_{\mathtt{NN}}(m_2, L_2, N_2, D) \end{array} \right\} \\ &\leq \inf \left\{ \sup_{\substack{\mathbf{x} \in [-R,R]^{m_1} \\ \mathbf{y} \in [-R,R]^{m_2}}} \left| f(\mathbf{x})^\top g(\mathbf{y}) - h(\mathbf{x}, \mathbf{y}) \right| \,\middle|\, \begin{array}{l} f \in \mathcal{F}_{\mathtt{NN}}(m_1, L_1, N_1, (2G+1)^{m_1}), \\ g \in \mathcal{F}_{\mathtt{NN}}(m_2, L_2, N_2, (2G+1)^{m_1}) \end{array} \right\} \\ &\leq \sup_{\substack{\mathbf{x} \in [-R,R]^{m_1} \\ \mathbf{y} \in [-R,R]^{m_2}}} \left| \tilde{v}(\mathbf{x})^\top \tilde{g}(\mathbf{y}) - h(\mathbf{x}, \mathbf{y}) \right| = \sup_{\substack{\mathbf{x} \in [-R,R]^{m_1} \\ \mathbf{y} \in [-R,R]^{m_2}}} |\tilde{q}(\mathbf{x}, \mathbf{y}) - h(\mathbf{x}, \mathbf{y})| \\ &\leq C_2 \cdot \left(\tfrac{1}{L_3^2 N_3^2} \right)^{\beta/m_2} + (2G+1)^{m_1} \cdot \left| m_1^2 \cdot 2^{-L_4} \right| \left| B + C_2 \cdot \left(\tfrac{1}{L_3^2 N_3^2} \right)^{\beta/m_2} \right| + B \cdot \left(\tfrac{R}{G} \right)^{\beta}.\end{aligned}$$

Since h is arbitrary, we obtain that

$$\begin{aligned}&\sup_{\substack{h: [-R,R]^{m_1+m_2} \to \mathbb{R} \\ \|h\|_{\beta, [-R,R]^{m_1+m_2}} \leq B}} \inf \left\{ \sup_{\substack{\mathbf{x} \in [-R,R]^{m_1} \\ \mathbf{y} \in [-R,R]^{m_2}}} \left| f(\mathbf{x})^\top g(\mathbf{y}) - h(\mathbf{x}, \mathbf{y}) \right| \,\middle|\, \begin{array}{l} f \in \mathcal{F}_{\mathtt{NN}}(m_1, L_1, N_1, D), \\ g \in \mathcal{F}_{\mathtt{NN}}(m_2, L_2, N_2, D) \end{array} \right\} \\ &\leq C_2 \cdot \left(\tfrac{1}{L_3^2 N_3^2} \right)^{\beta/m_2} + (2G+1)^{m_1} \cdot \left| m_1^2 \cdot 2^{-L_4} \right| \left| B + C_2 \cdot \left(\tfrac{1}{L_3^2 N_3^2} \right)^{\beta/m_2} \right| + B \cdot \left(\tfrac{R}{G} \right)^{\beta} \\ &\lesssim \left(\tfrac{1}{L_3^2 N_3^2} \right)^{\beta/m_2} + G^{m_1} \cdot 2^{-L_4} + \left(\tfrac{1}{G} \right)^{\beta} \lesssim \left(\tfrac{1}{L_3 N_3} \right)^{2\beta/m_2} + \left(\tfrac{1}{G} \right)^{\beta} + G^{m_1} \cdot 2^{-L_4} \\ &\lesssim \left| \epsilon^{\theta_4 - \min\{\theta_2, \theta_4, \theta_5\}} \cdot \epsilon^{\theta_3} \right|^{2\beta/m_2} + \left| \epsilon^{\min\{\theta_2, \theta_4, \theta_5\}} \right|^{\beta/m_1} \\ &\qquad + \epsilon^{-\min\{\theta_2, \theta_4, \theta_5\}} \cdot 2^{-\epsilon^{-\theta_1/2}} \\ &\lesssim \left| \epsilon^{\theta_3 + \theta_4 - \min\{\theta_2, \theta_4, \theta_5\}} \right|^{2\beta/m_2} + \left| \epsilon^{\min\{\theta_2, \theta_4, \theta_5\}} \right|^{\beta/m_1}.\end{aligned}$$

This proves the desired result. □

Lemma D.5. *Suppose $d_1 \in \mathbb{N}$, $d_2 \in \mathbb{N}$, $R \in (0, \infty)$, $\beta \in (0, 1]$, and $B \in (0, \infty)$. Fix arbitrary positive numbers $\theta_{1,1}$, $\theta_{1,2}$, $\theta_{2,1}$, $\theta_{2,2}$, θ_0. Denote $\alpha_1 := \theta_{1,1} + \theta_{1,2}$, $\alpha_2 := \theta_{2,1} + \theta_{2,2}$, and $\alpha_0 := \min\{\theta_0, \theta_{1,2}, \theta_{2,2}\}$. Then for sufficiently small $\epsilon \in (0, \infty)$ and $L_1 \asymp \epsilon^{-\theta_{1,1}}$, $N_1 \asymp \epsilon^{-\theta_{1,2}}$, $L_2 \asymp \epsilon^{-\theta_{2,1}}$, $N_2 \asymp \epsilon^{-\theta_{2,2}}$, $D \asymp \epsilon^{-\theta_0}$ there holds*

$$\begin{aligned}&\sup_{\substack{h: [-R,R]^{d_1+d_2} \to \mathbb{R} \\ \|h\|_{\beta, [-R,R]^{d_1+d_2}} \leq B}} \inf \left\{ \sup_{\substack{\mathbf{x} \in [-R,R]^{d_1} \\ \mathbf{y} \in [-R,R]^{d_2}}} \left| f(\mathbf{x})^\top g(\mathbf{y}) - h(\mathbf{x}, \mathbf{y}) \right| \,\middle|\, \begin{array}{l} f \in \mathcal{F}_{\mathtt{NN}}(d_1, L_1, N_1, D), \\ g \in \mathcal{F}_{\mathtt{NN}}(d_2, L_2, N_2, D) \end{array} \right\} \\ &\lesssim \min \left\{ \left| \epsilon^{\alpha_2 - \alpha_0} \right|^{\frac{2\beta}{d_2}} + \left| \epsilon^{\alpha_0} \right|^{\frac{\beta}{d_1}}, \left| \epsilon^{\alpha_1 - \alpha_0} \right|^{\frac{2\beta}{d_1}} + \left| \epsilon^{\alpha_0} \right|^{\frac{\beta}{d_2}} \right\}.\end{aligned}$$

Proof. Applying Lemma D.4 directly, we get

$$\sup_{\substack{h:[-R,R]^{d_1+d_2}\to\mathbb{R}\\ \|h\|_{\beta,[-R,R]^{d_1+d_2}}\le B}} \inf\left\{ \sup_{\substack{\mathbf{x}\in[-R,R]^{d_1}\\ \mathbf{y}\in[-R,R]^{d_2}}} \left|f(\mathbf{x})^\top g(\mathbf{y}) - h(\mathbf{x},\mathbf{y})\right| \;\middle|\; \begin{array}{l} f\in\mathcal{F}_{\texttt{NN}}(d_1,L_1,N_1,D)\\ g\in\mathcal{F}_{\texttt{NN}}(d_2,L_2,N_2,D)\end{array}\right\}$$
$$\lesssim \left|\epsilon^{\alpha_2-\alpha_0}\right|^{\frac{2\beta}{d_2}} + \left|\epsilon^{\alpha_0}\right|^{\frac{\beta}{d_1}},$$

and

$$\sup_{\substack{h:[-R,R]^{d_2+d_1}\to\mathbb{R}\\ \|h\|_{\beta,[-R,R]^{d_2+d_1}}\le B}} \inf\left\{ \sup_{\substack{\mathbf{x}\in[-R,R]^{d_2}\\ \mathbf{y}\in[-R,R]^{d_1}}} \left|f(\mathbf{x})^\top g(\mathbf{y}) - h(\mathbf{x},\mathbf{y})\right| \;\middle|\; \begin{array}{l} f\in\mathcal{F}_{\texttt{NN}}(d_2,L_2,N_2,D)\\ g\in\mathcal{F}_{\texttt{NN}}(d_1,L_1,N_1,D)\end{array}\right\}$$
$$\lesssim \left|\epsilon^{\alpha_1-\alpha_0}\right|^{\frac{2\beta}{d_1}} + \left|\epsilon^{\alpha_0}\right|^{\frac{\beta}{d_2}}.$$

Obviously,

$$\sup_{\substack{h:[-R,R]^{d_1+d_2}\to\mathbb{R}\\ \|h\|_{\beta,[-R,R]^{d_1+d_2}}\le B}} \inf\left\{ \sup_{\substack{\mathbf{x}\in[-R,R]^{d_1}\\ \mathbf{y}\in[-R,R]^{d_2}}} \left|f(\mathbf{x})^\top g(\mathbf{y}) - h(\mathbf{x},\mathbf{y})\right| \;\middle|\; \begin{array}{l} f\in\mathcal{F}_{\texttt{NN}}(d_1,L_1,N_1,D)\\ g\in\mathcal{F}_{\texttt{NN}}(d_2,L_2,N_2,D)\end{array}\right\}$$
$$= \sup_{\substack{h:[-R,R]^{d_2+d_1}\to\mathbb{R}\\ \|h\|_{\beta,[-R,R]^{d_2+d_1}}\le B}} \inf\left\{ \sup_{\substack{\mathbf{x}\in[-R,R]^{d_2}\\ \mathbf{y}\in[-R,R]^{d_1}}} \left|f(\mathbf{x})^\top g(\mathbf{y}) - h(\mathbf{x},\mathbf{y})\right| \;\middle|\; \begin{array}{l} f\in\mathcal{F}_{\texttt{NN}}(d_2,L_2,N_2,D)\\ g\in\mathcal{F}_{\texttt{NN}}(d_1,L_1,N_1,D)\end{array}\right\}.$$

Combining the above, the desired result immediately follows. □

Theorem D.6. *Let Assumption 5.3 hold. Fix arbitrary positive numbers $\theta_{1,1}$, $\theta_{1,2}$, $\theta_{2,1}$, $\theta_{2,2}$, θ_0. Denote $\alpha_1 := \theta_{1,1}+\theta_{1,2}$, $\alpha_2 := \theta_{2,1}+\theta_{2,2}$, and $\alpha_0 := \min\{\theta_0,\theta_{1,2},\theta_{2,2}\}$. Then for sufficiently small $\epsilon\in(0,\infty)$ and $L_1\asymp\epsilon^{-\theta_{1,1}}$, $N_1\asymp\epsilon^{-\theta_{1,2}}$, $L_2\asymp\epsilon^{-\theta_{2,1}}$, $N_2\asymp\epsilon^{-\theta_{2,2}}$, $D\asymp\epsilon^{-\theta_0}$ there holds*

$$\inf\left\{\mathcal{L}(\tilde{s})-\mathcal{L}^* \mid \tilde{s}\in\mathcal{F}_{\texttt{S}}(L_1,N_1,L_2,N_2,D)\right\}$$
$$\lesssim \min\left\{\left|\epsilon^{\alpha_2-\alpha_0}\right|^{\frac{4\beta}{d_2}} + \left|\epsilon^{\alpha_0}\right|^{\frac{2\beta}{d_1}}, \left|\epsilon^{\alpha_1-\alpha_0}\right|^{\frac{4\beta}{d_1}} + \left|\epsilon^{\alpha_0}\right|^{\frac{2\beta}{d_2}}\right\}.$$

Proof. According to Lemma D.3 and Assumption 5.3, there exists a function $h:[-R,R]^{d_1+d_2}\to\mathbb{R}$ such that

$$h(\mathbf{x},\mathbf{y}) = s^*(\mathbf{x},\mathbf{y}),\ \forall\,(\mathbf{x},\mathbf{y})\in\mathcal{X}\times\mathcal{Y}, \tag{54}$$

and $\|h\|_{\beta,[-R,R]^{d_1+d_2}} \le \|s^*\|_{\beta,\mathcal{X}\times\mathcal{Y}} \le B$. For any $f\in\mathcal{F}_{\texttt{NN}}(d_1,L_1,N_1,D)$ and any $g\in\mathcal{F}_{\texttt{NN}}(d_2,L_2,N_2,D)$, denote

$$U_{f,g}:\mathbb{R}^{d_1}\times\mathbb{R}^{d_2},\ (\mathbf{x},\mathbf{y})\mapsto f(\mathbf{x})^\top g(\mathbf{y}).$$

Obviously, $U_{f,g}\in\mathcal{F}_{\texttt{S}}(L_1,N_1,L_2,N_2,D)$. Then it follows from Lemma D.2 that

$$\inf\left\{\mathcal{L}(\tilde{s})-\mathcal{L}^* \mid \tilde{s}\in\mathcal{F}_{\texttt{S}}(L_1,N_1,L_2,N_2,D)\right\} \le \mathcal{L}(U_{f,g})-\mathcal{L}^*$$
$$\le \mathbb{E}_{\mathbf{x}}\left[\sup_{\mathbf{y}\in\mathcal{Y}}\tfrac{|U_{f,g}(\mathbf{x},\mathbf{y})-s^*(\mathbf{x},\mathbf{y})|^2}{2\tau}\right] \overset{\because(54)}{=\!=\!=} \mathbb{E}_{\mathbf{x}}\left[\sup_{\mathbf{y}\in\mathcal{Y}}\tfrac{|U_{f,g}(\mathbf{x},\mathbf{y})-h(\mathbf{x},\mathbf{y})|^2}{2\tau}\right]$$
$$\le \frac{1}{2\tau}\cdot\sup_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}}|U_{f,g}(\mathbf{x},\mathbf{y})-h(\mathbf{x},\mathbf{y})|^2 = \frac{1}{2\tau}\cdot\sup_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}}\left|f(\mathbf{x})^\top g(\mathbf{y})-h(\mathbf{x},\mathbf{y})\right|^2.$$

Since f and g are arbitrary, we can use Lemma D.5 to obtain that

$$\begin{aligned}
&\inf\left\{\mathcal{L}(\tilde{s})-\mathcal{L}^* \mid \tilde{s}\in\mathcal{F}_{\mathtt{S}}(L_1,N_1,L_2,N_2,D)\right\}\\
&\le \frac{1}{2\tau}\cdot\left|\inf\left\{\sup_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}}\left|f(\mathbf{x})^\top g(\mathbf{y})-h(\mathbf{x},\mathbf{y})\right| \;\middle|\; \begin{array}{l} f\in\mathcal{F}_{\mathtt{NN}}(d_1,L_1,N_1,D)\\ g\in\mathcal{F}_{\mathtt{NN}}(d_2,L_2,N_2,D)\end{array}\right\}\right|^2\\
&\le \frac{1}{2\tau}\cdot\sup_{\substack{u:[-R,R]^{d_1+d_2}\to\mathbb{R}\\ \|u\|_{\beta,[-R,R]^{d_1+d_2}}\le B}}\left|\inf\left\{\sup_{\substack{\mathbf{x}\in\mathcal{X}\\ \mathbf{y}\in\mathcal{Y}}}\left|f(\mathbf{x})^\top g(\mathbf{y})-u(\mathbf{x},\mathbf{y})\right| \;\middle|\; \begin{array}{l} f\in\mathcal{F}_{\mathtt{NN}}(d_1,L_1,N_1,D)\\ g\in\mathcal{F}_{\mathtt{NN}}(d_2,L_2,N_2,D)\end{array}\right\}\right|^2\\
&\lesssim \left|\min\left\{\left|\epsilon^{\alpha_2-\alpha_0}\right|^{\frac{2\beta}{d_2}}+\left|\epsilon^{\alpha_0}\right|^{\frac{\beta}{d_1}},\left|\epsilon^{\alpha_1-\alpha_0}\right|^{\frac{2\beta}{d_1}}+\left|\epsilon^{\alpha_0}\right|^{\frac{\beta}{d_2}}\right\}\right|^2\\
&\lesssim \min\left\{\left|\epsilon^{\alpha_2-\alpha_0}\right|^{\frac{4\beta}{d_2}}+\left|\epsilon^{\alpha_0}\right|^{\frac{2\beta}{d_1}},\left|\epsilon^{\alpha_1-\alpha_0}\right|^{\frac{4\beta}{d_1}}+\left|\epsilon^{\alpha_0}\right|^{\frac{2\beta}{d_2}}\right\}.
\end{aligned}$$

This proves the desired result. □

Corollary D.7. *Let Assumption 5.3 hold. Denote* $I:=\{i\in\{1,2\}\mid d_i=\max\{d_1,d_2\}\}$. *Fix arbitrary positive numbers* $\theta_{1,1}$, $\theta_{1,2}$, $\theta_{2,1}$, $\theta_{2,2}$ *that satisfy* $\theta_{1,2}\ge 1$, $\theta_{2,2}\ge 1$, *and*

$$\max\left\{\theta_{i,1}+\theta_{i,2}\mid i\in I\right\}\ge 1+\tfrac{1}{2}\cdot\tfrac{\max\{d_1,d_2\}}{\min\{d_1,d_2\}}. \tag{55}$$

Then for sufficiently small $\delta\in(0,\infty)$ *and* $L_1\asymp\delta^{-\theta_{1,1}}$, $N_1\asymp\delta^{-\theta_{1,2}}$, $L_2\asymp\delta^{-\theta_{2,1}}$, $N_2\asymp\delta^{-\theta_{2,2}}$, $D\asymp\delta^{-1}$ *there holds*

$$\inf\left\{\mathcal{L}(\tilde{s})-\mathcal{L}^* \mid \tilde{s}\in\mathcal{F}_{\mathtt{S}}(L_1,N_1,L_2,N_2,D)\right\}\lesssim\delta^{\frac{2\beta}{\min\{d_1,d_2\}}}.$$

Proof. From (55) we see there exists $\iota\in I$ such that $\theta_{\iota,1}+\theta_{\iota,2}\ge 1+\frac{1}{2}\cdot\frac{\max\{d_1,d_2\}}{\min\{d_1,d_2\}}=1+\frac{d_\iota}{2\cdot d_{3-\iota}}$. Taking $\theta_0=1$ in Theorem D.6, we obtain that

$$\begin{aligned}
&\inf\left\{\mathcal{L}(\tilde{s})-\mathcal{L}^* \mid \tilde{s}\in\mathcal{F}_{\mathtt{S}}(L_1,N_1,L_2,N_2,D)\right\}\\
&\lesssim \min\left\{\left|\delta^{\theta_{2,1}+\theta_{2,2}-1}\right|^{\frac{4\beta}{d_2}}+\left|\delta\right|^{\frac{2\beta}{d_1}},\left|\delta^{\theta_{1,1}+\theta_{1,2}-1}\right|^{\frac{4\beta}{d_1}}+\left|\delta\right|^{\frac{2\beta}{d_2}}\right\}\lesssim\left|\delta^{\theta_{\iota,1}+\theta_{\iota,2}-1}\right|^{\frac{4\beta}{d_\iota}}+\left|\delta\right|^{\frac{2\beta}{d_{3-\iota}}}\\
&\lesssim\left|\delta^{1+\frac{d_\iota}{2\cdot d_{3-\iota}}-1}\right|^{\frac{4\beta}{d_\iota}}+\left|\delta\right|^{\frac{2\beta}{d_{3-\iota}}}=2\cdot\left|\delta\right|^{\frac{2\beta}{d_{3-\iota}}}=2\cdot\left|\delta\right|^{\frac{2\beta}{\min\{d_1,d_2\}}}.
\end{aligned}$$

This proves the desired result. □

Proof of Theorem 5.4. Take $\theta_{1,1}:=\frac{d_1}{4d_2}$, $\theta_{1,2}:=1+\frac{d_1}{4d_2}$, $\theta_{2,1}:=\frac{d_2}{4d_1}$, $\theta_{2,2}:=1+\frac{d_2}{4d_1}$. Obviously, these constants $\theta_{1,1}$, $\theta_{1,2}$, $\theta_{2,1}$, $\theta_{2,2}$ satisfy the conditions of Corollary D.7. We can then take $\delta=\epsilon^{\frac{\min\{d_1,d_2\}}{2\beta}}$ in Corollary D.7 to obtain the desired result. □

E Proof of Proposition 5.6

First, we state a formal version of Proposition 5.6.

Proposition E.1 (Spectral structure under latent classes)**.** *Under Assumption 5.5, the PMI operator* $T:L_2(P_{\mathcal{X}})\to L_2(P_{\mathcal{Y}})$ *has rank at most* K. *Let*

$$\begin{aligned}
e_k(\mathbf{x})&=\frac{1}{\sqrt{\pi_k}}\mathbb{I}\{z_{\mathbf{x}}=k\},\quad k\in[K],\\
b_k(\mathbf{y})&=\log\frac{p(z=k|\mathbf{y})}{\pi_k},\quad k\in[K],
\end{aligned}$$

and define the $K\times K$ *matrix* M *with entries* $M_{kl}=\mathbb{E}_{\mathbf{y}\sim p_{\mathcal{Y}}}[b_k(\mathbf{y})b_l(\mathbf{y})]$. *Then the non-zero eigenvalues* $\{\lambda_i\}_{i=1}^K$ *of* T^*T *are exactly the eigenvalues of* M, *with corresponding orthonormal*

eigenfunctions

$$\psi_i(\mathbf{x}) = \sum_{k=1}^{K} v_k^{(i)}\, e_k(\mathbf{x}),$$

where $v^{(i)} \in \mathbb{R}^K$ are the eigenvectors of M. The eigenfunctions of TT^* are

$$\phi_i(\mathbf{y}) = \frac{1}{\sqrt{\lambda_i}} \sum_{k=1}^{K} \sqrt{\pi_k}\, v_k^{(i)}\, b_k(\mathbf{y}).$$

Consequently, the subspaces spanned by the eigenfunctions and by the (log)conditional probability coincide:

$$\operatorname{span}\{\psi_1, \dots, \psi_K\} = \operatorname{span}\{p(\cdot = k \mid \mathbf{x})\}_{k=1}^{K}, \operatorname{span}\{\phi_1, \dots, \phi_K\} = \operatorname{span}\{\log p(\cdot = k \mid \mathbf{y})\}_{k=1}^{K}.$$

There also holds

$$\log \frac{p(\mathbf{x}, \mathbf{y})}{p_{\mathcal{X}}(\mathbf{x}) p_{\mathcal{Y}}(\mathbf{y})} = \sum_{k=1}^{K} p(k \mid \mathbf{x}) \log \frac{p(z = k|\mathbf{y})}{\pi_k}.$$

Proof of Proposition E.1. According to Bayes rule, the pointwise mutual information simplifies to be

$$\begin{aligned} \log \frac{p(\mathbf{x}, \mathbf{y})}{p_{\mathcal{X}}(\mathbf{x}) p_{\mathcal{Y}}(\mathbf{y})} &= \log \sum_{k=1}^{K} \frac{p(z = k \mid \mathbf{x})\, p(z = k \mid \mathbf{y})}{\pi_k} = \log \frac{p(z_{\mathbf{x}} \mid \mathbf{y})}{\pi_{z_{\mathbf{x}}}}. \\ &= \sum_{k=1}^{K} \mathbb{I}\{z_{\mathbf{x}} = k\} \log \frac{p(z = k|\mathbf{y})}{\pi_k} = \sum_{k=1}^{K} \mathbb{I}\{\mathbf{x} \in C_k\}\, b_k(\mathbf{y}), \end{aligned} \tag{56}$$

where $C_k = \{\mathbf{x} : z_{\mathbf{x}} = k\}$ and $b_k(\mathbf{y}) = \log \frac{p(z=k|\mathbf{y})}{\pi_k}$.

Finite rank. For any $f \in L_2(P_{\mathcal{X}})$,

$$\begin{aligned} (Tf)(\mathbf{y}) &= \int_{\mathcal{X}} \log \frac{p(\mathbf{x}, \mathbf{y})}{p_{\mathcal{X}}(\mathbf{x}) p_{\mathcal{Y}}(\mathbf{y})} f(\mathbf{x}) p_{\mathcal{X}}(\mathbf{x})\, d\mathbf{x} \\ &= \sum_{k=1}^{K} b_k(\mathbf{y}) \int_{C_k} f(\mathbf{x}) p_{\mathcal{X}}(\mathbf{x})\, d\mathbf{x}. \end{aligned}$$

Hence $\operatorname{Im}(T) \subseteq \operatorname{span}\{b_1, \dots, b_K\}$, so $\operatorname{rank}(T) \le K$.

Singular functions of T. The sets $C_1, \dots, C_K$ form a partition of $\mathcal{X}$. The functions $\{\mathbb{I}\{\mathbf{x} \in C_k\}\}_{k=1}^{K}$ are orthogonal in $L_2(P_{\mathcal{X}})$ with norms $\|\mathbb{I}\{\mathbf{x} \in C_k\}\|^2_{L_2(P_{\mathcal{X}})} = P_{\mathcal{X}}(C_k) = \pi_k$. Thus

$$e_k(\mathbf{x}) = \frac{\mathbb{I}\{\mathbf{x} \in C_k\}}{\sqrt{\pi_k}}, \quad k = 1, \dots, K,$$

is an orthonormal basis of $\mathcal{H}_0 = \operatorname{span}\{\mathbb{I}\{\mathbf{x} \in C_k\}\} \subset L_2(P_{\mathcal{X}})$.

Let $\mathcal{H}_0^{\perp}$ be the orthogonal complement of $\mathcal{H}_0$ in $L_2(P_{\mathcal{X}})$. For any $f \in \mathcal{H}_0^{\perp}$ we have $\int_{C_k} f(\mathbf{x}) p_{\mathcal{X}}(\mathbf{x}) d\mathbf{x} = 0$ for all k, and consequently $Tf = 0$. Therefore T^*T vanishes on $\mathcal{H}_0^{\perp}$ and its non-zero eigenspace lies in $\mathcal{H}_0$.

Take $f = \sum_{k=1}^{K} \alpha_k e_k \in \mathcal{H}_0$. Then

$$\int_{C_k} f(\mathbf{x}) p_{\mathcal{X}}(\mathbf{x})\, d\mathbf{x} = \sum_{l=1}^{K} \alpha_l \frac{1}{\sqrt{\pi_l}} \int_{C_k} \mathbb{I}\{\mathbf{x} \in C_l\} p_{\mathcal{X}}(\mathbf{x})\, d\mathbf{x} = \alpha_k \sqrt{\pi_k}.$$

Hence

$$(Tf)(\mathbf{y}) = \sum_{k=1}^{K} \alpha_k \sqrt{\pi_k}\, b_k(\mathbf{y}).$$

Now compute T^*Tf. For any $g \in L_2(P_{\mathcal{Y}})$,

$$(T^*g)(\mathbf{x}) = \int_{\mathcal{Y}} \log \frac{p(\mathbf{x},\mathbf{y})}{p_{\mathcal{X}}(\mathbf{x})p_{\mathcal{Y}}(\mathbf{y})} g(\mathbf{y}) p_{\mathcal{Y}}(\mathbf{y})\, d\mathbf{y}.$$

Using (56), for $\mathbf{x} \in C_k$,

$$(T^*g)(\mathbf{x}) = \int_{\mathcal{Y}} b_k(\mathbf{y}) g(\mathbf{y}) p_{\mathcal{Y}}(\mathbf{y})\, d\mathbf{y}.$$

Applying this with $g = Tf$ yields, for $\mathbf{x} \in C_k$,

$$\begin{aligned}(T^*Tf)(\mathbf{x}) &= \int_{\mathcal{Y}} b_k(\mathbf{y}) \Big(\sum_{l=1}^{K} \alpha_l \sqrt{\pi_l}\, b_l(\mathbf{y}) \Big) p_{\mathcal{Y}}(\mathbf{y})\, d\mathbf{y} \\ &= \sum_{l=1}^{K} \alpha_l \sqrt{\pi_l}\, M_{kl},\end{aligned}$$

where $M_{kl} = \mathbb{E}_{\mathbf{y}\sim p_{\mathcal{Y}}}[b_k(\mathbf{y}) b_l(\mathbf{y})]$.

Since $\mathbb{I}\{\mathbf{x} \in C_k\} = \sqrt{\pi_k}\, e_k(\mathbf{x})$, we rewrite

$$\begin{aligned}(T^*Tf)(\mathbf{x}) &= \sum_{k=1}^{K} e_k(\mathbf{x}) \sqrt{\pi_k} \sum_{l=1}^{K} \alpha_l \sqrt{\pi_l}\, M_{kl} \\ &= \sum_{k=1}^{K} \Big(\sum_{l=1}^{K} A_{kl} \alpha_l \Big) e_k(\mathbf{x}),\end{aligned}$$

with $A_{kl} = \sqrt{\pi_k \pi_l}\, M_{kl}$. Thus, in the orthonormal basis $\{e_k\}$, the operator T^*T is represented by the matrix A. A is symmetric and positive semidefinite because

$$v^\top A v = \Big\| \sum_k v_k \sqrt{\pi_k}\, b_k \Big\|^2_{L_2(P_{\mathcal{Y}})} \geq 0.$$

Let $\lambda_1 \geq \cdots \geq \lambda_K > 0$ be the non-zero eigenvalues of A and $v^{(1)}, \ldots, v^{(K)} \in \mathbb{R}^K$ the corresponding orthonormal eigenvectors. Define

$$\psi_i(\mathbf{x}) = \sum_{k=1}^{K} v_k^{(i)} e_k(\mathbf{x}), \quad i = 1, \ldots, K.$$

Then $T^*T\psi_i = \lambda_i \psi_i$ and $\langle \psi_i, \psi_j \rangle_{p_{\mathcal{X}}} = \delta_{ij}$ because $\{e_k\}$ is orthonormal and $v^{(i)}$ are orthonormal vectors. According to Lemma 4.1, the right singular functions are obtained by

$$\phi_i(\mathbf{y}) = \frac{1}{\sqrt{\lambda_i}} (T\psi_i)(\mathbf{y}) = \frac{1}{\sqrt{\lambda_i}} \sum_{k=1}^{K} v_k^{(i)} \sqrt{\pi_k}\, b_k(\mathbf{y}).$$

Equivalence of spanned subspaces. From the definitions,

$$\operatorname{span}\{\psi_1, \ldots, \psi_K\} = \operatorname{span}\{e_1, \ldots, e_K\} = \operatorname{span}\{\mathbb{I}\{\mathbf{x} \in C_k\}\} = \operatorname{span}\{p(z = k \mid \mathbf{x})\}.$$

Similarly,

$$\operatorname{span}\{\phi_1, \ldots, \phi_K\} = \operatorname{span}\{b_k(\mathbf{y})\} = \operatorname{span}\Big\{ \log \frac{p(z=k|\mathbf{y})}{\pi_k} \Big\} = \operatorname{span}\{\log p(z=k|\mathbf{y})\},$$

because the class priors are constants per class. This completes the proof. □

E.1 Discussion on a high-probability posterior condition

In the deterministic case, the PMI function is

$$\mathrm{PMI}(\mathbf{x},\mathbf{y}) = \sum_{k=1}^{K} \mathbf{1}\{z_{\mathbf{x}} = k\} \log \frac{P(z=k \mid \mathbf{y})}{\pi_k} =: h_K(\mathbf{x},\mathbf{y}),$$

so a rank-K score represents PMI exactly.

In high probability case, we suppose instead that $P(z = z_{\mathbf{x}} \mid \mathbf{x}) \geq 1-\delta$. In this case, the rank-K score $\tau h_K(\mathbf{x},\mathbf{y})$ can approximate the optimal scoring function s^* well, with an approximation error dependent on δ.

Let $\eta_k(\mathbf{x}) = p(z=k \mid \mathbf{x}), r_k(\mathbf{y}) = \frac{p(z=k|\mathbf{y})}{\pi_k}$. Conditional independence and Bayes' rule give $\frac{p(\mathbf{x},\mathbf{y})}{p_{\mathcal{X}}(\mathbf{x})p_{\mathcal{Y}}(\mathbf{y})} = \sum_{k=1}^{K} \eta_k(\mathbf{x}) r_k(\mathbf{y})$.

Define $N(\mathbf{x},\mathbf{y}) = \sum_{k \neq z_{\mathbf{x}}} \eta_k(\mathbf{x}) r_k(\mathbf{y}), \qquad \varepsilon_\delta(\mathbf{x},\mathbf{y}) = \log\left(1 + \frac{N(\mathbf{x},\mathbf{y})}{\eta_{z_{\mathbf{x}}}(\mathbf{x}) r_{z_{\mathbf{x}}}(\mathbf{y})}\right)$.

Then $\log \frac{p(\mathbf{x},\mathbf{y})}{p_{\mathcal{X}}(\mathbf{x})p_{\mathcal{Y}}(\mathbf{y})} = \log \eta_{z_{\mathbf{x}}}(\mathbf{x}) + \log r_{z_{\mathbf{x}}}(\mathbf{y}) + \varepsilon_\delta(\mathbf{x},\mathbf{y})$. Let $\pi_{\min} = \min_k \pi_k$. We have $r_k(\mathbf{y}) \leq \frac{1}{\pi_{\min}}, \qquad N(\mathbf{x},\mathbf{y}) \leq \frac{\delta}{\pi_{\min}}$. According to Assumption 4.2, there holds $\eta_{z_{\mathbf{x}}}(\mathbf{x}) r_{z_{\mathbf{x}}}(\mathbf{y}) \geq \frac{1}{\alpha} - \frac{\delta}{\pi_{\min}}$. Thus, whenever $\delta < \pi_{\min}/\alpha$, $0 \leq \varepsilon_\delta(\mathbf{x},\mathbf{y}) \leq -\log\left(1 - \frac{\alpha\delta}{\pi_{\min}}\right)$.

The term $\log \eta_{z_{\mathbf{x}}}(\mathbf{x})$ depends only on x and therefore does not affect the contrastive objective. Applying the same L_2 comparison argument as in Theorem 4.2 yields

$$\mathcal{L}(\tau h_K) - \mathcal{L}^* \lesssim \tau^2 \log^2\left(1 - \frac{\alpha\delta}{\pi_{\min}}\right).$$

Hence the PMI operator need not remain exactly rank K, but the rank-K score incurs only an additional δ-dependent approximation error. When $\delta = 0$, we recover the exact deterministic result.

F Proof of Theorem 5.7

Throughout this section, we will use $\mathrm{dist}(A_1, A_2)$ to denote the $\|\cdot\|_\infty$-distance between the sets A_1 and A_2 in $\mathbb{R}^d$, namely

$$\mathrm{dist}(A_1, A_2) := \inf\left\{\|\mathbf{x}-\mathbf{y}\|_\infty \;\middle|\; \mathbf{x} \in A_1, \mathbf{y} \in A_2\right\}.$$

Lemma F.1. *Let $d \in \mathbb{N}$, $\varnothing \neq A \subset \mathbb{R}^d$, $\varnothing \neq B \subset \mathbb{R}^d$, and $\delta \in (0,\infty)$. Suppose $\mathrm{dist}(A,B) \geq \delta$. Define*

$$f : \mathbb{R}^d \to \mathbb{R},\ \mathbf{x} \mapsto \frac{\mathrm{dist}(\{\mathbf{x}\}, B)}{\mathrm{dist}(\{\mathbf{x}\}, A) + \mathrm{dist}(\{\mathbf{x}\}, B)}.$$

Then f is well-defined, and $\|f\|_{1,\mathbb{R}^d} \leq \max\{1, 1/\delta\}$.

Proof. For any $\mathbf{x} \in \mathbb{R}^d$ and any $\epsilon \in (0,\infty)$, there exist $\mathbf{a} \in A$ and $\mathbf{b} \in B$ such that

$$\begin{aligned}&\mathrm{dist}(\{\mathbf{x}\}, A) + \mathrm{dist}(\{\mathbf{x}\}, B) > \|\mathbf{x}-\mathbf{a}\|_\infty - \epsilon + \|\mathbf{x}-\mathbf{b}\|_\infty - \epsilon \\ &\geq -2\epsilon + \|\mathbf{a}-\mathbf{b}\|_\infty \geq -2\epsilon + \mathrm{dist}(A,B) \geq -2\epsilon + \delta.\end{aligned}$$

Since $\mathbf{x}$ and ϵ are arbitrary, we get

$$\mathrm{dist}(\{\mathbf{x}\}, A) + \mathrm{dist}(\{\mathbf{x}\}, B) \geq \delta,\ \forall\, \mathbf{x} \in \mathbb{R}^d. \tag{57}$$

Therefore, the denominator in the definition of f is always positive, meaning that f is well defined. For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and any $\mathbf{a} \in A$, we have

$$\|\mathbf{x}-\mathbf{a}\|_\infty \leq \|\mathbf{x}-\mathbf{y}\|_\infty + \|\mathbf{y}-\mathbf{a}\|_\infty.$$

Taking the infimum over $\mathbf{a} \in A$ on both sides of the above inequality gives

$$\operatorname{dist}(\{\mathbf{x}\}, A) \leq \|\mathbf{x}-\mathbf{y}\|_\infty + \operatorname{dist}(\{\mathbf{y}\}, A),$$

meaning that

$$\operatorname{dist}(\{\mathbf{x}\}, A) - \operatorname{dist}(\{\mathbf{y}\}, A) \leq \|\mathbf{x}-\mathbf{y}\|_\infty. \tag{58}$$

Similarly, we have

$$\operatorname{dist}(\{\mathbf{y}\}, A) - \operatorname{dist}(\{\mathbf{x}\}, A) \leq \|\mathbf{y}-\mathbf{x}\|_\infty. \tag{59}$$

Combining (58) and (59), we obtain $|\operatorname{dist}(\{\mathbf{y}\}, A) - \operatorname{dist}(\{\mathbf{x}\}, A)| \leq \|\mathbf{y}-\mathbf{x}\|_\infty$. Since $\mathbf{x}, \mathbf{y}$ are arbitrary, we obtain

$$|\operatorname{dist}(\{\mathbf{y}\}, A) - \operatorname{dist}(\{\mathbf{x}\}, A)| \leq \|\mathbf{y}-\mathbf{x}\|_\infty, \ \forall\, \mathbf{x} \in \mathbb{R}^d, \mathbf{y} \in \mathbb{R}^d. \tag{60}$$

Similarly, we have

$$|\operatorname{dist}(\{\mathbf{y}\}, B) - \operatorname{dist}(\{\mathbf{x}\}, B)| \leq \|\mathbf{y}-\mathbf{x}\|_\infty, \ \forall\, \mathbf{x} \in \mathbb{R}^d, \mathbf{y} \in \mathbb{R}^d. \tag{61}$$

Obviously,

$$0 \leq f(\mathbf{x}) \leq 1 = \sup_{\mathbf{z}\in\mathbb{R}^d} f(\mathbf{z}), \ \forall\, \mathbf{x} \in \mathbb{R}^d. \tag{62}$$

Define

$$F : \{(w,z) \in [0,\infty) \times [0,\infty) \mid w+z \geq \delta\} \to [0,\infty), (x,y) \mapsto x/(x+y).$$

Then we have that

$$\begin{aligned}
&|F(x,y) - F(u,v)| = \left|\frac{x}{x+y} - \frac{u}{u+v}\right| = \left|\frac{xv-yu}{(u+v)\cdot(x+y)}\right| = \left|\frac{v\cdot(x-u)+u\cdot(v-y)}{(u+v)\cdot(x+y)}\right| \\
&\leq \left|\frac{v\cdot(x-u)}{(u+v)\cdot(x+y)}\right| + \left|\frac{u\cdot(v-y)}{(u+v)\cdot(x+y)}\right| \leq \left|\frac{v}{u+v}\right| \cdot \frac{|x-u|}{\delta} + \left|\frac{u}{u+v}\right| \cdot \frac{|v-y|}{\delta} \\
&\leq \left|\frac{v}{u+v}\right| \cdot \frac{\max\{|x-u|, |v-y|\}}{\delta} + \left|\frac{u}{u+v}\right| \cdot \frac{\max\{|x-u|, |v-y|\}}{\delta} \\
&= \frac{\max\{|x-u|, |v-y|\}}{\delta}, \ \forall\, (x,y),(u,v) \in \{(w,z) \in [0,\infty) \times [0,\infty) \mid w+z \geq \delta\},
\end{aligned}$$

which, together with (57), (60), (61), (62), yields that

$$\begin{aligned}
&|f(\mathbf{x}) - f(\mathbf{y})| = |F(\operatorname{dist}(\{\mathbf{x}\}, B), \operatorname{dist}(\{\mathbf{x}\}, A)) - F(\operatorname{dist}(\{\mathbf{y}\}, B), \operatorname{dist}(\{\mathbf{y}\}, A))| \\
&\leq \frac{\max\{|\operatorname{dist}(\{\mathbf{x}\}, B) - \operatorname{dist}(\{\mathbf{y}\}, B)|, |\operatorname{dist}(\{\mathbf{x}\}, A) - \operatorname{dist}(\{\mathbf{y}\}, A)|\}}{\delta} \\
&\leq \frac{\max\{\|\mathbf{x}-\mathbf{y}\|_\infty, \|\mathbf{x}-\mathbf{y}\|_\infty\}}{\delta} = \frac{\|\mathbf{x}-\mathbf{y}\|_\infty}{\delta}, \ \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|f\|_{1,\mathbb{R}^d} &= \max\left\{\sup_{\mathbf{z}\in\mathbb{R}^d} |f(\mathbf{z})|, \sup_{\mathbb{R}^d \ni \mathbf{x}\neq\mathbf{y}\in\mathbb{R}^d} \frac{|f(\mathbf{x}) - f(\mathbf{y})|}{\|\mathbf{x}-\mathbf{y}\|_\infty}\right\} \\
&\leq \max\left\{\sup_{\mathbf{z}\in\mathbb{R}^d} |f(\mathbf{z})|, \sup_{\mathbb{R}^d \ni \mathbf{x}\neq\mathbf{y}\in\mathbb{R}^d} \frac{\|\mathbf{x}-\mathbf{y}\|_\infty/\delta}{\|\mathbf{x}-\mathbf{y}\|_\infty}\right\} \\
&= \max\left\{\sup_{\mathbf{z}\in\mathbb{R}^d} |f(\mathbf{z})|, 1/\delta\right\} \overset{\because (62)}{=\!=} \max\{1, 1/\delta\}.
\end{aligned}$$

This proves the desired result. □

Lemma F.2. *Suppose $R \in (0,\infty)$, $B \in (0,\infty)$, $\beta \in (0,1]$, $d \in \mathbb{N}$. Then for sufficiently small $\epsilon \in (0,\infty)$ and $L_1 \gtrsim \log\frac{1}{\epsilon}$, $N_1 \gtrsim \log\frac{1}{\epsilon}$ there holds*

$$\sup_{\substack{h:[-R,R]^d\to\mathbb{R} \\ \|h\|_{\beta,[-R,R]^d} \leq B}} \inf\left\{\sup_{\mathbf{y}\in[-R,R]^d} |f(\mathbf{y}) - h(\mathbf{y})| \,\middle|\, f \in \mathcal{F}_{\mathtt{NN}}(d, L_1, N_1, 1)\right\} \lesssim \frac{1}{\log^{4\beta/d}(1/\epsilon)}.$$

Proof. According to Lemma D.1, there exist constants $C_1 = C_1(d) \in \mathbb{N}$ and $C_2 = C_2(d, R, \beta, B) \in (0, \infty)$ such that for any $L \in \mathbb{N}$, $N \in \mathbb{N}$ and any $h : [-R, R]^d \to \mathbb{R}$ with $\|h\|_{\beta,[-R,R]^d} \leq B$ there exists a neural network

$$f \in \mathcal{F}_{\mathtt{NN}}(d, C_1 L, C_1 N, 1)$$

such that

$$\sup_{\mathbf{y} \in [-R,R]^d} |f(\mathbf{y}) - h(\mathbf{y})| \leq C_2 \cdot \left(\tfrac{1}{L^2 N^2}\right)^{\beta/d}.$$

Take $N_3 := \lfloor N_1 / C_1 \rfloor$ and

$$L_3 := \sup \left\{ L \in \mathbb{N} \cap [2, \infty) \mid C_1 L \leq L_1 \right\}.$$

Then we have that $C_1 L_3 \leq L_1$, $C_1 N_3 \leq N_1$, $N_3 \asymp N_1 \gtrsim \log \frac{1}{\epsilon}$, and $L_3 \asymp L_1 \gtrsim \log \frac{1}{\epsilon}$. Therefore, for any $h : [-R, R]^d \to \mathbb{R}$ with $\|h\|_{\beta,[-R,R]^d} \leq B$ there exists a neural network

$$f \in \mathcal{F}_{\mathtt{NN}}(d, C_1 L_3, C_1 N_3, 1) \subset \mathcal{F}_{\mathtt{NN}}(d, L_1, N_1, 1)$$

such that

$$\sup_{\mathbf{y} \in [-R,R]^d} |f(\mathbf{y}) - h(\mathbf{y})| \leq C_2 \cdot \left(\tfrac{1}{L_3^2 N_3^2}\right)^{\beta/d}.$$

Since h is arbitrary, we obtain that

$$\sup_{\substack{h:[-R,R]^d \to \mathbb{R} \\ \|h\|_{\beta,[-R,R]^d} \leq B}} \inf \left\{ \sup_{\mathbf{y} \in [-R,R]^d} |f(\mathbf{y}) - h(\mathbf{y})| \;\middle|\; f \in \mathcal{F}_{\mathtt{NN}}(d, L_1, N_1, 1) \right\} \leq C_2 \cdot \left(\tfrac{1}{L_3^2 N_3^2}\right)^{\beta/d}$$

$$\lesssim \left(\tfrac{1}{L_3^2 N_3^2}\right)^{\beta/d} \lesssim \left(\tfrac{1}{L_1^2 N_1^2}\right)^{\beta/d} \lesssim \left(\tfrac{1}{\log^4(1/\epsilon)}\right)^{\beta/d} = \tfrac{1}{\log^{4\beta/d}(1/\epsilon)}.$$

This proves the desired result. □

Lemma F.3. *Suppose $R \in (0, \infty)$, $B \in (0, \infty)$, $\beta \in (0, 1]$, $d \in \mathbb{N}$. Fix an arbitrary positive number θ. Then for sufficiently small $\epsilon \in (0, \infty)$ and $L_2 \gtrsim \log \frac{1}{\epsilon}$, $N_2 \gtrsim \log \frac{1}{\epsilon}$ with $L_2 \cdot N_2 \asymp \epsilon^{-\theta}$ there holds*

$$\sup_{\substack{h:[-R,R]^d \to \mathbb{R} \\ \|h\|_{\beta,[-R,R]^d} \leq B}} \inf \left\{ \sup_{\mathbf{y} \in [-R,R]^d} |f(\mathbf{y}) - h(\mathbf{y})| \;\middle|\; f \in \mathcal{F}_{\mathtt{NN}}(d, L_2, N_2, 1) \right\} \lesssim \epsilon^{2\theta\beta/d}.$$

Proof. According to Lemma D.1, there exist constants $C_1 = C_1(d) \in \mathbb{N}$ and $C_2 = C_2(d, R, \beta, B) \in (0, \infty)$ such that for any $L \in \mathbb{N}$, $N \in \mathbb{N}$ and any $h : [-R, R]^d \to \mathbb{R}$ with $\|h\|_{\beta,[-R,R]^d} \leq B$ there exists a neural network

$$f \in \mathcal{F}_{\mathtt{NN}}(d, C_1 L, C_1 N, 1)$$

such that

$$\sup_{\mathbf{y} \in [-R,R]^d} |f(\mathbf{y}) - h(\mathbf{y})| \leq C_2 \cdot \left(\tfrac{1}{L^2 N^2}\right)^{\beta/d}.$$

Take $N_3 := \lfloor N_2 / C_1 \rfloor$ and

$$L_3 := \sup \left\{ L \in \mathbb{N} \cap [2, \infty) \mid C_1 L \leq L_2 \right\}.$$

Then we have that $C_1 L_3 \leq L_2$, $C_1 N_3 \leq N_2$, $N_3 \asymp N_2 \gtrsim \log \frac{1}{\epsilon}$, and $L_3 \asymp L_2 \gtrsim \log \frac{1}{\epsilon}$. Therefore, for any $h : [-R, R]^d \to \mathbb{R}$ with $\|h\|_{\beta,[-R,R]^d} \leq B$ there exists a neural network

$$f \in \mathcal{F}_{\mathtt{NN}}(d, C_1 L_3, C_1 N_3, 1) \subset \mathcal{F}_{\mathtt{NN}}(d, L_2, N_2, 1)$$

such that

$$\sup_{\mathbf{y}\in[-R,R]^d} |f(\mathbf{y}) - h(\mathbf{y})| \leq C_2 \cdot \left(\tfrac{1}{L_3^2 N_3^2}\right)^{\beta/d}.$$

Since h is arbitrary, we obtain that

$$\sup_{\substack{h:[-R,R]^d\to\mathbb{R} \\ \|h\|_{\beta,[-R,R]^d}\leq B}} \inf\left\{ \sup_{\mathbf{y}\in[-R,R]^d} |f(\mathbf{y}) - h(\mathbf{y})| \;\middle|\; f \in \mathcal{F}_{\texttt{NN}}(d, L_2, N_2, 1) \right\}$$
$$\leq C_2 \cdot \left(\tfrac{1}{L_3^2 N_3^2}\right)^{\beta/d} \lesssim \left(\tfrac{1}{L_2^2 N_2^2}\right)^{\beta/d} \asymp \left(\tfrac{1}{\epsilon^{-2\theta}}\right)^{\beta/d} = \epsilon^{2\theta\beta/d}.$$

This completes the proof. □

Proof of Theorem 5.7. For simplicity, we denote $A_k := \{\mathbf{x} \in \mathcal{X} \mid p(z = k \mid \mathbf{x}) = 1\}$, and $g_k(\mathbf{y}) := \tau \cdot \log \frac{p(z=k|\mathbf{y})}{\pi_k}$. According to Assumption 5.5, $\{A_k\}_{k=1}^K$ is a partition of $\mathcal{X}$. It follows from Proposition E.1 that

$$s^*(\mathbf{x},\mathbf{y}) := \tau \log \frac{p(\mathbf{y}|\mathbf{x})}{p_{\mathcal{Y}}(\mathbf{y})} = \tau \log \frac{p(\mathbf{x},\mathbf{y})}{p_{\mathcal{X}}(\mathbf{x}) p_{\mathcal{Y}}(\mathbf{y})} = \tau \cdot \sum_{k=1}^K p(z=k \mid \mathbf{x}) \log \frac{p(z=k \mid \mathbf{y})}{\pi_k}$$
$$= \tau \cdot \sum_{k=1}^K \mathbb{I}\{\mathbf{x} \in A_k\} \cdot \log \frac{p(z=k \mid \mathbf{y})}{\pi_k} = \sum_{k=1}^K g_k(\mathbf{y}) \cdot \mathbb{I}\{\mathbf{x} \in A_k\}, \ \forall\, (\mathbf{x},\mathbf{y}) \in \mathcal{X} \times \mathcal{Y}.$$

Take $J := \{k \in \{1,2,\ldots,K\} \mid A_k \neq \varnothing\}$, and $\mathcal{G} := \{g_k \mid k \in J\}$. For each $f \in \mathcal{G}$, denote

$$B_f := \bigcup_{k\in J, g_k = f} A_k.$$

Then we have that $\{B_f\}_{f\in\mathcal{G}}$ is a partition of $\mathcal{X}$, and

$$s^*(\mathbf{x},\mathbf{y}) = \sum_{k=1}^K g_k(\mathbf{y}) \cdot \mathbb{I}\{\mathbf{x} \in A_k\} = \sum_{k\in J} g_k(\mathbf{y}) \cdot \mathbb{I}\{\mathbf{x} \in A_k\}$$
$$= \sum_{f\in\mathcal{G}} \sum_{\substack{k\in J \\ g_k = f}} g_k(\mathbf{y}) \cdot \mathbb{I}\{\mathbf{x} \in A_k\} = \sum_{f\in\mathcal{G}} \sum_{\substack{k\in J \\ g_k = f}} f(\mathbf{y}) \cdot \mathbb{I}\{\mathbf{x} \in A_k\}$$
$$= \sum_{f\in\mathcal{G}} f(\mathbf{y}) \cdot \mathbb{I}\{\mathbf{x} \in B_f\}, \ \forall\, (\mathbf{x},\mathbf{y}) \in \mathcal{X} \times \mathcal{Y}.$$

For each $\mathcal{G} \ni f_1 \neq f_2 \in \mathcal{G}$, there exists $\mathbf{y}_0 \in \mathcal{Y}$ such that $f_1(\mathbf{y}_0) \neq f_2(\mathbf{y}_0)$. Then for any $\mathbf{x}_1 \in B_{f_1}$ and any $\mathbf{x}_2 \in B_{f_2}$, we have that

$$|f_1(\mathbf{y}_0) - f_2(\mathbf{y}_0)| = |s^*(\mathbf{x}_1,\mathbf{y}_0) - s^*(\mathbf{x}_2,\mathbf{y}_0)| \leq \|s^*\|_{\beta,\mathcal{X}\times\mathcal{Y}} \cdot \|(\mathbf{x}_1,\mathbf{y}_0) - (\mathbf{x}_2,\mathbf{y}_0)\|_\infty^\beta$$
$$\leq B \cdot \|(\mathbf{x}_1,\mathbf{y}_0) - (\mathbf{x}_2,\mathbf{y}_0)\|_\infty^\beta = B \cdot \|\mathbf{x}_1 - \mathbf{x}_2\|_\infty^\beta,$$

meaning that $\|\mathbf{x}_1 - \mathbf{x}_2\|_\infty \geq \left(\frac{|f_1(\mathbf{y}_0) - f_2(\mathbf{y}_0)|}{B}\right)^{1/\beta}$. Since $\mathbf{x}_1$, $\mathbf{x}_2$ are arbitrary, we obtain that

$$\operatorname{dist}(B_{f_1}, B_{f_2}) = \inf_{\mathbf{x}_1\in B_{f_1}, \mathbf{x}_2 \in B_{f_2}} \|\mathbf{x}_1 - \mathbf{x}_2\|_\infty \geq \left(\frac{|f_1(\mathbf{y}_0) - f_2(\mathbf{y}_0)|}{B}\right)^{1/\beta} > 0.$$

Since f_1 and f_2 are arbitrary and $|\mathcal{G}| \leq K < \infty$, we get

$$\delta := \inf_{f_1\in\mathcal{G}} \operatorname{dist}(B_{f_1}, \mathcal{X} \setminus B_{f_1}) = \inf_{f_1\in\mathcal{G}} \operatorname{dist}(B_{f_1}, \textstyle\bigcup_{f_2\in\mathcal{G}\setminus\{f_1\}} B_{f_2})$$
$$= \inf_{f_1\in\mathcal{G}} \inf_{f_2\in\mathcal{G}\setminus\{f_1\}} \operatorname{dist}(B_{f_1}, B_{f_2}) = \inf_{\mathcal{G}\ni f_1 \neq f_2 \in \mathcal{G}} \operatorname{dist}(B_{f_1}, B_{f_2}) > 0.$$

For any $f \in \mathcal{G}$, define

$$U_f : \mathbb{R}^{d_1} \to [0,1],\ \mathbf{x} \mapsto \frac{\operatorname{dist}(\{\mathbf{x}\}, \mathcal{X} \setminus B_f)}{\operatorname{dist}(\{\mathbf{x}\}, B_f) + \operatorname{dist}(\{\mathbf{x}\}, \mathcal{X} \setminus B_f)}$$

if $\mathcal{X} \setminus B_f \neq \varnothing$; define $U_f(\mathbf{x}) :\equiv 1$ if $\mathcal{X} \setminus B_f = \varnothing$. According to Lemma F.1,

$$\|U_f\|_{1,[-R,R]^{d_1}} \leq \|U_f\|_{1,\mathbb{R}^{d_1}} \leq \max\{1, 1/\delta\},\ \forall\ f \in \mathcal{G}.$$

Note that $L_1 - 1 \gtrsim L_1 \gtrsim \log \frac{1}{\epsilon}$ and $\lfloor N_1/K \rfloor \gtrsim N_1 \gtrsim \log \frac{1}{\epsilon}$. Then it follows from Lemma F.2 that

$$\begin{aligned}
&\sup_{\substack{h:[-R,R]^{d_1} \to \mathbb{R} \\ \|h\|_{1,[-R,R]^{d_1}} \leq \max\left\{1, \frac{1}{\delta}\right\}}} \inf \left\{ \sup_{\mathbf{y} \in [-R,R]^{d_1}} |f(\mathbf{y}) - h(\mathbf{y})| \,\middle|\, f \in \mathcal{F}_{\texttt{NN}}(d_1, L_1 - 1, \lfloor N_1/K \rfloor, 1) \right\} \\
&\lesssim \frac{1}{\log^{4/d_1}(1/\epsilon)} \to 0.
\end{aligned}$$

Therefore, in what follows, we may assume that ϵ is sufficiently small so that

$$\begin{aligned}
&\sup_{\substack{h:[-R,R]^{d_1} \to \mathbb{R} \\ \|h\|_{1,[-R,R]^{d_1}} \leq \max\left\{1, \frac{1}{\delta}\right\}}} \inf \left\{ \sup_{\mathbf{y} \in [-R,R]^{d_1}} |f(\mathbf{y}) - h(\mathbf{y})| \,\middle|\, f \in \mathcal{F}_{\texttt{NN}}(d_1, L_1 - 1, \lfloor N_1/K \rfloor, 1) \right\} \\
&< 1/9 < 3 \leq \lfloor N_1/K \rfloor.
\end{aligned}$$

Thus, for any $f \in \mathcal{G}$, there exists

$$\widetilde{U}_f \in \mathcal{F}_{\texttt{NN}}(d_1, L_1 - 1, \lfloor N_1/K \rfloor, 1) \tag{63}$$

such that

$$\sup_{\mathbf{x} \in [-R,R]^{d_1}} \left| U_f(\mathbf{x}) - \widetilde{U}_f(\mathbf{x}) \right| \leq 1/9.$$

Therefore,

$$\widetilde{U}_f(\mathbf{x}) \geq U_f(\mathbf{x}) - 1/9 = 1 - 1/9 = 8/9,\ \forall\ \mathbf{x} \in B_f,\ \forall\ f \in \mathcal{G},$$

and

$$\widetilde{U}_f(\mathbf{x}) \leq U_f(\mathbf{x}) + 1/9 = 0 + 1/9 = 1/9,\ \forall\ \mathbf{x} \in \mathcal{X} \setminus B_f,\ \forall\ f \in \mathcal{G}.$$

We then define $T(t) := 3\sigma(t - 1/3) - 3\sigma(t - 2/3)$ and $\widetilde{H}_f := T \circ \widetilde{U}_f$. Obviously,

$$T(x) = 0 < 1 = T(y),\ \forall\ x \in (-\infty, 1/3],\ \forall\ y \in [2/3, \infty).$$

Hence

$$\widetilde{H}_f(\mathbf{x}) = T(\widetilde{U}_f(\mathbf{x})) \xlongequal{\because \widetilde{U}_f(\mathbf{x}) \geq 8/9} 1,\ \forall\ \mathbf{x} \in B_f,\ \forall\ f \in \mathcal{G},$$

and

$$\widetilde{H}_f(\mathbf{x}) = T(\widetilde{U}_f(\mathbf{x})) \xlongequal{\because \widetilde{U}_f(\mathbf{x}) \leq 1/9} 0,\ \forall\ \mathbf{x} \in \mathcal{X} \setminus B_f,\ \forall\ f \in \mathcal{G},$$

meaning that

$$\widetilde{H}_f(\mathbf{x}) = \mathbb{I}\{\mathbf{x} \in B_f\},\ \forall\ \mathbf{x} \in \mathcal{X},\ \forall\ f \in \mathcal{G}. \tag{64}$$

Note that

$$T \in \mathcal{F}_{\texttt{NN}}(1, 1, 2, 1) \subset \mathcal{F}_{\texttt{NN}}(1, 1, \lfloor N_1/K \rfloor, 1),$$

which, together with (63), yields

$$\widetilde{H}_f = T \circ \widetilde{U}_f \in \mathcal{F}_{\texttt{NN}}(d_1, L_1, \lfloor N_1/K \rfloor, 1),\ \forall\ f \in \mathcal{G}. \tag{65}$$

For any $f \in \mathcal{G}$, there exist $k \in J$ and $\mathbf{x}_0 \in A_k$ such that

$$s^*(\mathbf{x}_0, \mathbf{y}) = \sum_{j=1}^{K} g_j(\mathbf{y}) \cdot \mathbb{I}\{\mathbf{x}_0 \in A_j\} = g_k(\mathbf{y}) = f(\mathbf{y}),\ \forall\, y \in \mathcal{Y},$$

which, together with Assumption 5.3, yields

$$\|f\|_{\beta,\mathcal{Y}} = \|s^*(\mathbf{x}_0, \cdot)\|_{\beta,\mathcal{Y}} \le \|s^*\|_{\beta,\mathcal{X}\times\mathcal{Y}} \le B.$$

Thus

$$\|f\|_{\beta,\mathcal{Y}} \le B,\ \forall\, f \in \mathcal{G}.$$

Then it follows from Lemma D.3 that for any $f \in \mathcal{G}$ there exists $Q_f : [-R,R]^{d_2} \to \mathbb{R}$ such that $\|Q_f\|_{\beta,[-R,R]^{d_2}} \le \|f\|_{\beta,\mathcal{Y}} \le B$, and

$$f(\mathbf{y}) = Q_f(\mathbf{y}),\ \forall\, \mathbf{y} \in \mathcal{Y}. \tag{66}$$

We then take

$$\widetilde{Q}_f \in \mathcal{F}_{\texttt{NN}}(d_2, L_2, \lfloor N_2/K \rfloor, 1) \tag{67}$$

such that

$$\begin{aligned}
&\sup_{\mathbf{y}\in[-R,R]^{d_2}} \left| Q_f(\mathbf{y}) - \widetilde{Q}_f(\mathbf{y}) \right| \\
&\le \inf \left\{ \sup_{\mathbf{y}\in[-R,R]^{d_2}} |Q_f(\mathbf{y}) - u(\mathbf{y})| \,\middle|\, u \in \mathcal{F}_{\texttt{NN}}(d_2, L_2, \lfloor N_2/K \rfloor, 1) \right\} + \epsilon.
\end{aligned}$$

Define $\widetilde{S}(\mathbf{x},\mathbf{y}) := \sum_{f\in\mathcal{G}} \widetilde{H}_f(\mathbf{x}) \cdot \widetilde{Q}_f(\mathbf{y})$. Combining (65) and (67), we obtain that

$$\begin{aligned}
&\widetilde{S} \in \mathcal{F}_{\texttt{S}}(L_1, \lfloor N_1/K \rfloor \cdot |\mathcal{G}|, L_2, \lfloor N_2/K \rfloor \cdot |\mathcal{G}|, |\mathcal{G}|) \\
&\subset \mathcal{F}_{\texttt{S}}(L_1, \lfloor N_1/K \rfloor \cdot K, L_2, \lfloor N_2/K \rfloor \cdot K, K) \subset \mathcal{F}_{\texttt{S}}(L_1, N_1, L_2, N_2, K).
\end{aligned}$$

Then it follows from Lemma D.2 that

$$\begin{aligned}
&\inf\{\mathcal{L}(\tilde{s}) - \mathcal{L}^* \mid \tilde{s} \in \mathcal{F}_{\texttt{S}}(L_1, N_1, L_2, N_2, K)\} \le \mathcal{L}(\widetilde{S}) - \mathcal{L}^* \\
&\le \mathbb{E}_{\mathbf{x}} \left[\sup_{\mathbf{y}\in\mathcal{Y}} \frac{\left|\widetilde{S}(\mathbf{x},\mathbf{y}) - s^*(\mathbf{x},\mathbf{y})\right|^2}{2\tau} \right] \le \sup_{(\mathbf{x},\mathbf{y})\in\mathcal{X}\times\mathcal{Y}} \left|\widetilde{S}(\mathbf{x},\mathbf{y}) - s^*(\mathbf{x},\mathbf{y})\right|^2 \cdot \frac{1}{2\tau} \\
&= \sup_{(\mathbf{x},\mathbf{y})\in\mathcal{X}\times\mathcal{Y}} \left| \sum_{f\in\mathcal{G}} \widetilde{H}_f(\mathbf{x}) \cdot \widetilde{Q}_f(\mathbf{y}) - \sum_{f\in\mathcal{G}} f(\mathbf{y}) \cdot \mathbb{I}\{\mathbf{x} \in B_f\} \right|^2 \cdot \frac{1}{2\tau} \\
&\overset{\because (64)}{=\!=} \sup_{(\mathbf{x},\mathbf{y})\in\mathcal{X}\times\mathcal{Y}} \left| \sum_{f\in\mathcal{G}} \mathbb{I}\{\mathbf{x} \in B_f\} \cdot \widetilde{Q}_f(\mathbf{y}) - \sum_{f\in\mathcal{G}} f(\mathbf{y}) \cdot \mathbb{I}\{\mathbf{x} \in B_f\} \right|^2 \cdot \frac{1}{2\tau} \\
&= \sup_{f\in\mathcal{G}} \sup_{\mathbf{y}\in Y} \left|\widetilde{Q}_f(\mathbf{y}) - f(\mathbf{y})\right|^2 \cdot \frac{1}{2\tau} \overset{\because (66)}{=\!=} \sup_{f\in\mathcal{G}} \sup_{\mathbf{y}\in Y} \left|\widetilde{Q}_f(\mathbf{y}) - Q_f(\mathbf{y})\right|^2 \cdot \frac{1}{2\tau} \\
&\le \sup_{f\in\mathcal{G}} \sup_{\mathbf{y}\in[-R,R]^{d_2}} \left|\widetilde{Q}_f(\mathbf{y}) - Q_f(\mathbf{y})\right|^2 \cdot \frac{1}{2\tau}.
\end{aligned}$$

Thus

$$\begin{aligned}
&\inf\{\mathcal{L}(\tilde{s}) - \mathcal{L}^* \mid \tilde{s} \in \mathcal{F}_{\texttt{S}}(L_1, N_1, L_2, N_2, K)\} \le \sup_{f\in\mathcal{G}} \sup_{\mathbf{y}\in[-R,R]^{d_2}} \left|\widetilde{Q}_f(\mathbf{y}) - Q_f(\mathbf{y})\right|^2 \cdot \frac{1}{2\tau} \\
&\le \frac{1}{2\tau} \cdot \sup_{f\in\mathcal{G}} \left| \inf \left\{ \sup_{\mathbf{y}\in[-R,R]^{d_2}} |Q_f(\mathbf{y}) - u(\mathbf{y})| \,\middle|\, u \in \mathcal{F}_{\texttt{NN}}(d_2, L_2, \lfloor N_2/K \rfloor, 1) \right\} + \epsilon \right|^2 \\
&\le \frac{\epsilon^2}{\tau} + \frac{1}{\tau} \cdot \sup_{f\in\mathcal{G}} \left| \inf \left\{ \sup_{\mathbf{y}\in[-R,R]^{d_2}} |Q_f(\mathbf{y}) - u(\mathbf{y})| \,\middle|\, u \in \mathcal{F}_{\texttt{NN}}(d_2, L_2, \lfloor N_2/K \rfloor, 1) \right\} \right|^2 \\
&\le \frac{\epsilon^2}{\tau} + \frac{1}{\tau} \cdot \sup_{\substack{h:[-R,R]^{d_2}\to\mathbb{R} \\ \|h\|_{\beta,[-R,R]^{d_2}} \le B}} \left| \inf \left\{ \sup_{\mathbf{y}\in[-R,R]^{d_2}} |h(\mathbf{y}) - u(\mathbf{y})| \,\middle|\, u \in \mathcal{F}_{\texttt{NN}}(d_2, L_2, \lfloor N_2/K \rfloor, 1) \right\} \right|^2.
\end{aligned}$$

Hence

$$\begin{aligned}
&\inf\left\{\mathcal{L}(\tilde{s})-\mathcal{L}^* \mid \tilde{s}\in\mathcal{F}_{\mathtt{S}}(L_1,N_1,L_2,N_2,K)\right\}\\
&\leq \frac{\epsilon^2}{\tau}+\frac{1}{\tau}\cdot\sup_{\substack{h:[-R,R]^{d_2}\to\mathbb{R}\\ \|h\|_{\beta,[-R,R]^{d_2}}\leq B}}\left|\inf\left\{\sup_{\mathbf{y}\in[-R,R]^{d_2}}|h(\mathbf{y})-u(\mathbf{y})|\;\middle|\; u\in\mathcal{F}_{\mathtt{NN}}(d_2,L_2,\lfloor N_2/K\rfloor,1)\right\}\right|^2.
\end{aligned}$$

Note that $L_2\cdot\lfloor N_2/K\rfloor\asymp L_2\cdot N_2\asymp\epsilon^{-\frac{d_2}{4\beta}}$. Therefore, we can use Lemma F.3 to obtain that

$$\begin{aligned}
&\inf\left\{\mathcal{L}(\tilde{s})-\mathcal{L}^* \mid \tilde{s}\in\mathcal{F}_{\mathtt{S}}(L_1,N_1,L_2,N_2,K)\right\}\\
&\lesssim \epsilon^2+\sup_{\substack{h:[-R,R]^{d_2}\to\mathbb{R}\\ \|h\|_{\beta,[-R,R]^{d_2}}\leq B}}\left|\inf\left\{\sup_{\mathbf{y}\in[-R,R]^{d_2}}|h(\mathbf{y})-u(\mathbf{y})|\;\middle|\; u\in\mathcal{F}_{\mathtt{NN}}(d_2,L_2,\lfloor N_2/K\rfloor,1)\right\}\right|^2\\
&\lesssim \epsilon^2+\left|\epsilon^{1/2}\right|^2\lesssim\epsilon.
\end{aligned}$$

This completes the proof. □

G Pseudo-Dimension of $\mathtt{PD}_{\mathbb{R}^{d_1+d_2}}\left(\mathcal{F}_{\mathtt{S}}(L_1,N_1,L_2,N_2,D)\right)$

Let Ω be a nonempty set and $\mathcal{G}$ be a class of real-valued functions whose domains contain Ω. The pseudo-dimension of $\mathcal{G}$ on Ω is defined as

$$\mathtt{PD}_{\Omega}(\mathcal{G}):=\sup\left\{m\in\mathbb{N}\mid \mathbf{G}_{\Omega,\mathcal{G}}(m)=2^m\right\},$$

where $\mathbf{G}_{\Omega,\mathcal{G}}(m)$ is the so-called growth function of $\mathcal{G}$ on Ω, defined as

$$\begin{aligned}
&\mathbf{G}_{\Omega,\mathcal{G}}(m)\\
&:=\max_{\substack{\mathbf{x}_1,\ldots,\mathbf{x}_m\in\Omega\\ t_1,\ldots,t_m\in\mathbb{R}}}\left|\left\{\left(\mathbb{I}\left\{f(\mathbf{x}_1)\geq t_1\right\},\mathbb{I}\left\{f(\mathbf{x}_2)\geq t_2\right\},\ldots,\mathbb{I}\left\{f(\mathbf{x}_m)\geq t_m\right\}\right)\;\middle|\; f\in\mathcal{G}\right\}\right|.
\end{aligned}$$

Theorem G.1. *For any $L_1\in\mathbb{N}$, $L_2\in\mathbb{N}$, $N_1\in\mathbb{N}\cap(1,\infty)$, $N_2\in\mathbb{N}\in(1,\infty)$, $D\in\mathbb{N}$, there holds*

$$\begin{aligned}
&\mathtt{PD}_{\mathbb{R}^{d_1+d_2}}\left(\mathcal{F}_{\mathtt{S}}(L_1,N_1,L_2,N_2,D)\right)\\
&\leq 32(L_1^2N_1^2+L_1N_1d_1+N_1D+L_2^2N_2^2+L_2N_2d_2+N_2D)\log(L_1L_2+L_1N_1+L_2N_2).
\end{aligned}$$

Proof. For $j=1,2$, we write

$$\begin{aligned}
\mathbf{f}_j(\mathbf{x},\mathbf{w}_j)&=\left(f_{j,1}(\mathbf{x},\mathbf{w}_j),f_{j,2}(\mathbf{x},\mathbf{w}_j),\ldots,f_{j,D}(\mathbf{x},\mathbf{w}_j)\right)^\top\\
&=\boldsymbol{W}_{L_j}\sigma_{\mathbf{v}_{L_j}}\boldsymbol{W}_{L_j-1}\sigma_{\mathbf{v}_{L_j-1}}\cdots\boldsymbol{W}_2\sigma_{\mathbf{v}_2}\boldsymbol{W}_1\sigma_{\mathbf{v}_1}\boldsymbol{W}_0\mathbf{x},
\end{aligned}$$

where $\mathbf{v}_i\in\mathbb{R}^{N_j}$ $(1\leq i\leq L_j)$, $\boldsymbol{W}_i\in\mathbb{R}^{N_j\times N_j}$ $(1\leq i\leq L_j-1)$, $\boldsymbol{W}_0\in\mathbb{R}^{N_j\times d_j}$, $\boldsymbol{W}_{L_j}\in\mathbb{R}^{D\times N_j}$, and $\mathbf{w}_j$ is the vector whose components consist of all the components of $\boldsymbol{W}_0,\ldots,\boldsymbol{W}_{L_j},\mathbf{v}_1,\ldots\mathbf{v}_{L_j}$. For $\iota\in\{1,2,\ldots,L_j\}$, we use $k_{j,\iota}$ to denote the total number of entries in $\boldsymbol{W}_0,\ldots,\boldsymbol{W}_{\iota-1},\mathbf{v}_1,\ldots\mathbf{v}_\iota$. Then let k_{j,L_j+1} denote the total number of entries in $\boldsymbol{W}_0,\ldots,\boldsymbol{W}_{L_j},\mathbf{v}_1,\ldots\mathbf{v}_\iota$. Thus $k_{j,L_j+1}=(L_j+d_j+D)\cdot N_j+(L_j-1)\cdot N_j^2$, and $\mathbf{w}_j\in\mathbb{R}^{k_{j,L_j+1}}$. For simplicity, we denote $K_j:=k_{j,L_j+1}$, $\Omega:=\mathbb{R}^{d_1+d_2}$, and

$$S(\mathbf{x},\mathbf{y},\mathbf{w}_1,\mathbf{w}_2):=\mathbf{f}_1(\mathbf{x},\mathbf{w}_1)^\top\mathbf{f}_2(\mathbf{y},\mathbf{w}_2)=\sum_{i=1}^{D}f_{1,i}(\mathbf{x},\mathbf{w}_1)f_{2,i}(\mathbf{y},\mathbf{w}_2).$$

Obviously,

$$\mathcal{G}:=\left\{(\mathbf{x},\mathbf{y})\mapsto S(\mathbf{x},\mathbf{y},\mathbf{w}_1,\mathbf{w}_2)\;\middle|\;\mathbf{w}_1\in\mathbb{R}^{K_1},\mathbf{w}_2\in\mathbb{R}^{K_2}\right\}=\mathcal{F}_{\mathtt{S}}(L_1,N_1,L_2,N_2,D),$$

meaning that

$$\mathtt{PD}_{\Omega}\left(\mathcal{F}_{\mathtt{S}}(L_1,N_1,L_2,N_2,D)\right)=\mathtt{PD}_{\Omega}\left(\mathcal{G}\right)=\sup\left\{m\in\mathbb{N}\mid\mathbf{G}_{\Omega,\mathcal{G}}(m)=2^m\right\}.\tag{68}$$

We next establish upper bounds for the growth function $\mathbf{G}_{\Omega,\mathcal{G}}(m)$.

Let $m \in \mathbb{N}$, $\mathbf{x}_{1,1}, \ldots, \mathbf{x}_{1,m} \in \mathbb{R}^{d_1}$, $\mathbf{x}_{2,1}, \ldots, \mathbf{x}_{2,m} \in \mathbb{R}^{d_2}$, and $t_1, \ldots, t_m \in \mathbb{R}$ be arbitrary. By the arguments used in the proof of Theorem 7 in Bartlett et al. [2019], there exist a partition $\mathcal{U}_1$ of $\mathbb{R}^{K_1}$ and a partition $\mathcal{U}_2$ of $\mathcal{R}^{K_2}$, such that

$$
\begin{aligned}
|\mathcal{U}_j| &\leq \prod_{l=1}^{L_j} 2 \cdot \left(\frac{2\mathrm{e}mN_j \cdot 1 \cdot (1 + (l-1) \cdot 1^{l-1})}{k_{j,l}} \right)^{k_{j,l}} \\
&\leq \prod_{l=1}^{L_j} \left(\frac{12mN_j \cdot L_j}{k_{j,l}} \right)^{k_{j,l}}, \ \forall\, j \in \{1,2\},
\end{aligned}
$$

and that for any $j \in \{1,2\}$, $U \in \mathcal{U}_j$, $r \in \{1,2,\ldots,m\}$, $i \in \{1,2,\ldots,D\}$, the function

$$
U \ni \mathbf{w} \mapsto f_{j,i}(\mathbf{x}_{j,r}, \mathbf{w}) \in \mathbb{R}
$$

coincides with a polynomial with total degree no more than $(1 + L_j)$. Therefore, for any $U_1 \in \mathcal{U}_1$, $U_2 \in \mathcal{U}_2$, $r \in \{1,2,\ldots,m\}$, the function

$$
U_1 \times U_2 \ni (\mathbf{w}_1, \mathbf{w}_2) \mapsto S(\mathbf{x}_{1,r}, \mathbf{x}_{2,r}, \mathbf{w}_1, \mathbf{w}_2) \in \mathbb{R}
$$

coincides with a polynomial, which we denote by Q_{r,U_1,U_2}, with total degree no more than $(1 + L_1)(1 + L_2)$. We can then apply Lemma 17 of Bartlett et al. [2019] to obtain that

$$
\begin{aligned}
&\left| \left\{ (\mathbb{I}\{H(\mathbf{x}_{1,1}, \mathbf{x}_{2,1}) \geq t_1\}, \ldots, \mathbb{I}\{H(\mathbf{x}_{1,m}, \mathbf{x}_{2,m}) \geq t_m\}) \;\middle|\; H \in \mathcal{G} \right\} \right| \\
&= \left| \bigcup_{\substack{\mathbf{w}_1 \in \mathbb{R}^{K_1} \\ \mathbf{w}_2 \in \mathbb{R}^{K_2}}} \{(\mathbb{I}\{S(\mathbf{x}_{1,1}, \mathbf{x}_{2,1}, \mathbf{w}_1, \mathbf{w}_2) \geq t_1\}, \ldots, \mathbb{I}\{S(\mathbf{x}_{1,m}, \mathbf{x}_{2,m}, \mathbf{w}_1, \mathbf{w}_2) \geq t_m\})\} \right| \\
&\leq \sum_{\substack{U_1 \in \mathcal{U}_1 \\ U_2 \in \mathcal{U}_2}} \left| \bigcup_{\substack{\mathbf{w}_1 \in U_1 \\ \mathbf{w}_2 \in U_2}} \{(\mathbb{I}\{S(\mathbf{x}_{1,1}, \mathbf{x}_{2,1}, \mathbf{w}_1, \mathbf{w}_2) \geq t_1\}, \ldots, \mathbb{I}\{S(\mathbf{x}_{1,m}, \mathbf{x}_{2,m}, \mathbf{w}_1, \mathbf{w}_2) \geq t_m\})\} \right| \\
&= \sum_{\substack{U_1 \in \mathcal{U}_1 \\ U_2 \in \mathcal{U}_2}} \left| \bigcup_{\substack{\mathbf{w}_1 \in U_1 \\ \mathbf{w}_2 \in U_2}} \{(\mathbb{I}\{t_1 - Q_{1,U_1,U_2}(\mathbf{w}_1, \mathbf{w}_2) \leq 0\}, \ldots, \mathbb{I}\{t_m - Q_{m,U_1,V_1}(\mathbf{w}_1, \mathbf{w}_2) \leq 0\})\} \right| \\
&= \sum_{\substack{U_1 \in \mathcal{U}_1 \\ U_2 \in \mathcal{U}_2}} \left| \bigcup_{\substack{\mathbf{w}_1 \in U_1 \\ \mathbf{w}_2 \in U_2}} \{(\mathbb{I}\{t_1 - Q_{1,U_1,U_2}(\mathbf{w}_1, \mathbf{w}_2) > 0\}, \ldots, \mathbb{I}\{t_m - Q_{m,U_1,V_1}(\mathbf{w}_1, \mathbf{w}_2) > 0\})\} \right| \\
&\leq \sum_{\substack{U_1 \in \mathcal{U}_1 \\ U_2 \in \mathcal{U}_2}} 2 \cdot (2\mathrm{e}m(1 + L_1)(1 + L_2)/(K_1 + K_2))^{K_1 + K_2}.
\end{aligned}
$$

Thus

$$
\begin{aligned}
&\left| \left\{ (\mathbb{I}\{H(\mathbf{x}_{1,1}, \mathbf{x}_{2,1}) \geq t_1\}, \ldots, \mathbb{I}\{H(\mathbf{x}_{1,m}, \mathbf{x}_{2,m}) \geq t_m\}) \;\middle|\; H \in \mathcal{G} \right\} \right| \\
&\leq \sum_{\substack{U_1 \in \mathcal{U}_1 \\ U_2 \in \mathcal{U}_2}} 2 \cdot (2\mathrm{e}m(1 + L_1)(1 + L_2)/(K_1 + K_2))^{K_1 + K_2} \\
&= |\mathcal{U}_1|\,|\mathcal{U}_2| 2 \cdot (2\mathrm{e}m(1 + L_1)(1 + L_2)/(K_1 + K_2))^{K_1 + K_2} \\
&\leq \left(\frac{48mL_1L_2}{K_1 + K_2} \right)^{K_1 + K_2} \prod_{j=1}^{2} \prod_{l=1}^{L_j} \left(\frac{12mN_j \cdot L_j}{k_{j,l}} \right)^{k_{j,l}}.
\end{aligned}
$$

Since $\mathbf{x}_{1,1},\dots,\mathbf{x}_{1,m}\in\mathbb{R}^{d_1}$, $\mathbf{x}_{2,1},\dots,\mathbf{x}_{2,m}\in\mathbb{R}^{d_2}$, and $t_1,\dots,t_m\in\mathbb{R}$ are arbitrary, we get

$$\begin{aligned}
&\mathbf{G}_{\Omega,\mathcal{G}}(m)\\
&=\max_{\substack{\mathbf{x}_{1,1},\dots,\mathbf{x}_{1,m}\in\mathbb{R}^{d_1}\\ \mathbf{x}_{2,1},\dots,\mathbf{x}_{2,m}\in\mathbb{R}^{d_2}\\ t_1,\dots,t_m\in\mathbb{R}}}\left|\left\{\left(\mathbb{I}\left\{H(\mathbf{x}_{1,1},\mathbf{x}_{2,1})\geq t_1\right\},\dots,\mathbb{I}\left\{H(\mathbf{x}_{1,m},\mathbf{x}_{2,m})\geq t_m\right\}\right)\;\middle|\; H\in\mathcal{G}\right\}\right|\\
&\leq\left(\frac{48mL_1L_2}{K_1+K_2}\right)^{K_1+K_2}\prod_{j=1}^{2}\prod_{l=1}^{L_j}\left(\frac{12mN_j\cdot L_j}{k_{j,l}}\right)^{k_{j,l}}.
\end{aligned}$$

Using the weighted AM-GM inequality, we obtain

$$\begin{aligned}
\mathbf{G}_{\Omega,\mathcal{G}}(m)&\leq\left(\frac{48mL_1L_2}{K_1+K_2}\right)^{K_1+K_2}\prod_{j=1}^{2}\prod_{l=1}^{L_j}\left(\frac{12mN_j\cdot L_j}{k_{j,l}}\right)^{k_{j,l}}\\
&\leq\left(\frac{48mL_1L_2+12mN_1L_1^2+12mN_2L_2^2}{K_1+K_2+\sum_{j=1}^{2}\sum_{l=1}^{L_j}k_{j,l}}\right)^{K_1+K_2+\sum_{j=1}^{2}\sum_{l=1}^{L_j}k_{j,l}}.
\end{aligned}$$

Denote $A=48L_1L_2+12N_1L_1^2+12N_2L_2^2$, $W=K_1+K_2+\sum_{j=1}^{2}\sum_{l=1}^{L_j}k_{j,l}$. The we have that

$$\mathbf{G}_{\Omega,\mathcal{G}}(m)\leq\left(\frac{Am}{W}\right)^{W}.$$

Since m is arbitrary, we obtain that $\mathtt{PD}_{\Omega}(\mathcal{G})=\sup\{m\in\mathbb{N}\mid\mathbf{G}_{\Omega,\mathcal{G}}(m)=2^m\}<\infty$, and

$$2^{\mathtt{PD}_{\Omega}(\mathcal{G})}=\mathbf{G}_{\Omega,\mathcal{G}}(\mathtt{PD}_{\Omega}(\mathcal{G}))\leq\left(\frac{A\cdot\mathtt{PD}_{\Omega}(\mathcal{G})}{W}\right)^{W}.$$

We can then apply Lemma 18 of Bartlett et al. [2019] to derive from the above inequality that

$$\begin{aligned}
&\mathtt{PD}_{\Omega}(\mathcal{G})\leq W\log_2(2A\log_2 A)\leq 4W\log A\leq(K_1+K_2+L_1k_{1,L_1}+L_2k_{2,L_2})\cdot 4\log A\\
&\leq(K_1+K_2+L_1k_{1,L_1}+L_2k_{2,L_2})\cdot 16\log(L_1L_2+L_1N_1+L_2N_2)\\
&=\left|\sum_{j=1}^{2}((L_j+1)\left((L_j-1)N_j^2+L_jN_j+N_jd_j\right)+N_jD)\right|\cdot 16\log(L_1L_2+L_1N_1+L_2N_2)\\
&\leq\left|\sum_{j=1}^{2}(L_j^2N_j^2+L_jN_jd_j+N_jD)\right|\cdot 32\log(L_1L_2+L_1N_1+L_2N_2)\\
&=32(L_1^2N_1^2+L_1N_1d_1+N_1D+L_2^2N_2^2+L_2N_2d_2+N_2D)\log(L_1L_2+L_1N_1+L_2N_2),
\end{aligned}$$

which, together with (68), proves the desired inequality. □

H Proof of Theorem 6.2

Throughout this section, we write $r(\mathbf{x},\mathbf{y}):=p(\mathbf{y}\mid\mathbf{x})/p_{\mathcal{Y}}(\mathbf{y})$ for simplicity. We first recall the following technical lemma from Li et al. [2026, Theorem E.14]. It controls the excess risk of ranking through the misalignment of the order induced by scoring function and density ratio as well as the gap of density ratio.

Lemma H.1. *Let Assumption 4.2 hold. For any measurable scoring function $s:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$, define*

$$A_s:=\left\{(\mathbf{x},\mathbf{y},\mathbf{y}')\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}:\left(r(\mathbf{x},\mathbf{y})-r(\mathbf{x},\mathbf{y}')\right)\left(s(\mathbf{x},\mathbf{y})-s(\mathbf{x},\mathbf{y}')\right)\leq 0\right\}.$$

Then

$$\mathcal{E}^*-\mathcal{E}(s)\leq\frac{1}{2}\mathbb{E}_{\mathbf{x},\mathbf{y},\mathbf{y}'}\left[\mathbb{I}\{(\mathbf{x},\mathbf{y},\mathbf{y}')\in A_s\}|r(\mathbf{x},\mathbf{y})-r(\mathbf{x},\mathbf{y}')|\right].\tag{69}$$

We next state the Hellinger–KL inequality used to control the difference between the square roots of two densities by their KL divergence.

Lemma H.2 (Hellinger–KL inequality)**.** *Let P and Q be probability measures on the same measurable space, with densities p and q with respect to a common measure ν. Suppose that $P \ll Q$. Define the squared Hellinger distance by*

$$H^2(P,Q) := \int (\sqrt{p} - \sqrt{q})^2 \, d\nu.$$

Then

$$H^2(P,Q) \le \mathrm{KL}(P \,\|\, Q), \tag{70}$$

where $\mathrm{KL}(P \,\|\, Q) = \int p \log(p/q) \, d\nu$.

Proof. Suppose $\mathrm{KL}(P \,\|\, Q) < +\infty$, then $q > 0$ holds P-almost everywhere. We apply the elementary inequality $\log u \le u - 1$ for $u > 0$ to $u = \sqrt{q/p}$ on $\{p > 0\}$, gives

$$p \log \frac{p}{q} = -2p \log \sqrt{\frac{q}{p}} \ge 2p - 2\sqrt{pq}.$$

Integrating over $\{p > 0\}$ and using $\int p \, d\nu = \int q \, d\nu = 1$, we obtain

$$\mathrm{KL}(P \,\|\, Q) \ge 2 - 2 \int \sqrt{pq} \, d\nu = \int (\sqrt{p} - \sqrt{q})^2 \, d\nu = H^2(P,Q).$$

This completes the proof. □

Proof of Theorem 6.2. Define $Z_s(\mathbf{x}) := \mathbb{E}_{\mathbf{y} \sim P_{\mathcal{Y}}} \exp(s(\mathbf{x},\mathbf{y})/\tau)$ and

$$a_s(\mathbf{x},\mathbf{y}) := \frac{\exp(s(\mathbf{x},\mathbf{y})/\tau)}{Z_s(\mathbf{x})}, \qquad dQ_{s,\mathbf{x}}(\mathbf{y}) := a_s(\mathbf{x},\mathbf{y}) dP_{\mathcal{Y}}(\mathbf{y}).$$

Then $Q_{s,\mathbf{x}}$ is a probability measure on $\mathcal{Y}$. We further have

$$\mathcal{L}(s) - \mathcal{L}^* = \tau \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{X}}} \left[\mathrm{KL}\big(P_{\mathcal{Y}|\mathbf{x}} \,\|\, Q_{s,\mathbf{x}}\big) \right], \tag{71}$$

where $P_{\mathcal{Y}|\mathbf{x}}$ denotes the conditional distribution on $\mathcal{Y}$ given $\mathbf{x}$. Define $h_s(\mathbf{x},\mathbf{y}) := \sqrt{r(\mathbf{x},\mathbf{y})} - \sqrt{a_s(\mathbf{x},\mathbf{y})}$. Applying Lemma H.2 to $P_{\mathcal{Y}|\mathbf{x}}$ and $Q_{s,\mathbf{x}}$, whose densities with respect to $P_{\mathcal{Y}}$ are $r(\mathbf{x},\cdot)$ and $a_s(\mathbf{x},\cdot)$, and then taking expectation over $\mathbf{x}$, we obtain

$$\mathbb{E}_{\mathbf{x},\mathbf{y}} \left[h_s(\mathbf{x},\mathbf{y})^2 \right] = \mathbb{E}_{\mathbf{x}} \int \left(\sqrt{\frac{dP_{\mathcal{Y}|\mathbf{x}}(\mathbf{y})}{dP_{\mathcal{Y}}(\mathbf{y})}} - \sqrt{\frac{dQ_{s,\mathbf{x}}(\mathbf{y})}{dP_{\mathcal{Y}}(\mathbf{y})}} \right)^2 dP_{\mathcal{Y}}(\mathbf{y}) \le \frac{\mathcal{L}(s) - \mathcal{L}^*}{\tau}. \tag{72}$$

Fix $(\mathbf{x},\mathbf{y},\mathbf{y}') \in A_s$. Without loss of generality, we assume that $r(\mathbf{x},\mathbf{y}) \ge r(\mathbf{x},\mathbf{y}')$. If $r(\mathbf{x},\mathbf{y}) > r(\mathbf{x},\mathbf{y}')$, the definition of A_s gives $s(\mathbf{x},\mathbf{y}) \le s(\mathbf{x},\mathbf{y}')$, and hence $a_s(\mathbf{x},\mathbf{y}) \le a_s(\mathbf{x},\mathbf{y}')$. It follows that

$$0 \le \sqrt{r(\mathbf{x},\mathbf{y})} - \sqrt{r(\mathbf{x},\mathbf{y}')} \le |h_s(\mathbf{x},\mathbf{y})| + |h_s(\mathbf{x},\mathbf{y}')|.$$

Using $r(\mathbf{x},\mathbf{y}), r(\mathbf{x},\mathbf{y}') \le \alpha$, we obtain

$$\begin{aligned} |r(\mathbf{x},\mathbf{y}) - r(\mathbf{x},\mathbf{y}')|^2 &= \big(\sqrt{r(\mathbf{x},\mathbf{y})} + \sqrt{r(\mathbf{x},\mathbf{y}')}\big)^2 \big(\sqrt{r(\mathbf{x},\mathbf{y})} - \sqrt{r(\mathbf{x},\mathbf{y}')}\big)^2 \\ &\le 4\alpha \big(|h_s(\mathbf{x},\mathbf{y})| + |h_s(\mathbf{x},\mathbf{y}')|\big)^2 \\ &\le 8\alpha \big(h_s(\mathbf{x},\mathbf{y})^2 + h_s(\mathbf{x},\mathbf{y}')^2\big). \end{aligned}$$

The above inequality naturally holds for $r(\mathbf{x},\mathbf{y}) = r(\mathbf{x},\mathbf{y}')$. Taking expectations and applying (72), we have

$$\begin{aligned} &\mathbb{E}_{\mathbf{x},\mathbf{y},\mathbf{y}'} \left[\mathbb{I}\{(\mathbf{x},\mathbf{y},\mathbf{y}') \in A_s\} |r(\mathbf{x},\mathbf{y}) - r(\mathbf{x},\mathbf{y}')|^2 \right] \le \mathbb{E}_{\mathbf{x},\mathbf{y},\mathbf{y}'} \left[8\alpha \big(h_s(\mathbf{x},\mathbf{y})^2 + h_s(\mathbf{x},\mathbf{y}')^2\big) \right] \\ =&16\alpha \mathbb{E}_{\mathbf{x},\mathbf{y}} \left[h_s(\mathbf{x},\mathbf{y})^2 \right] \le \frac{16\alpha}{\tau} \left(\mathcal{L}(s) - \mathcal{L}^* \right). \end{aligned} \tag{73}$$

For any $t > 0$, Lemma H.1, (73), and Assumption 6.1 imply

$$\begin{aligned}
\mathcal{E}^* - \mathcal{E}(s) &\le \frac{1}{2}\mathbb{E}_{\mathbf{x},\mathbf{y},\mathbf{y}'}[\mathbb{I}\{(\mathbf{x},\mathbf{y},\mathbf{y}') \in A_s\}|r(\mathbf{x},\mathbf{y}) - r(\mathbf{x},\mathbf{y}')|\mathbb{I}\{|r(\mathbf{x},\mathbf{y}) - r(\mathbf{x},\mathbf{y}')| > t\}] \\
&\quad + \frac{1}{2}\mathbb{E}_{\mathbf{x},\mathbf{y},\mathbf{y}'}[\mathbb{I}\{(\mathbf{x},\mathbf{y},\mathbf{y}') \in A_s\}|r(\mathbf{x},\mathbf{y}) - r(\mathbf{x},\mathbf{y}')|\mathbb{I}\{|r(\mathbf{x},\mathbf{y}) - r(\mathbf{x},\mathbf{y}')| \le t\}] \\
&\le \frac{1}{2t}\mathbb{E}_{\mathbf{x},\mathbf{y},\mathbf{y}'}[\mathbb{I}\{(\mathbf{x},\mathbf{y},\mathbf{y}') \in A_s\}|r(\mathbf{x},\mathbf{y}) - r(\mathbf{x},\mathbf{y}')|^2] + \frac{t}{2}\mathbb{P}(|r(\mathbf{x},\mathbf{y}) - r(\mathbf{x},\mathbf{y}')| \le t) \\
&\le \frac{8\alpha(\mathcal{L}(s) - \mathcal{L}^*)}{\tau t} + \frac{ct^{\mu+1}}{2}.
\end{aligned}$$

Taking $t = (16\alpha(\mathcal{L}(s) - \mathcal{L}^*)/(c\tau))^{1/(\mu+2)}$ gives

$$\mathcal{E}^* - \mathcal{E}(s) \le c^{\frac{1}{\mu+2}} \left(\frac{16\alpha(\mathcal{L}(s) - \mathcal{L}^*)}{\tau}\right)^{\frac{\mu+1}{\mu+2}}.$$

This completes the proof. □

I Numerical Experiments

In this section, we We conduct experiments on Gaussian mixture models to examine three aspects of our theory. First, a balanced Gaussian-mixture model with infinite rank illustrates the prediction in Section 4 that approximation error decreases as the representation dimension increases. Second, we train neural networks on imbalanced Gaussian mixture models to examine the representation structure and dimension dependence described in Section 5.2. Finally, we verify the density-ratio upper bound in Assumption 4.2 and Tsybakov noise condition in Assumption 6.1.

I.1 Approximation error in a Gaussian mixture model

Setup. We first study spectral decay and approximation error as the representation dimension D increases in a balanced binary Gaussian mixture model with an infinite-rank PMI operator. Let z be uniform on $\{-1, +1\}$, with conditionally independent views $\mathbf{x} \mid z \sim \mathcal{N}(z\mu_x, \sigma_x^2)$ and $\mathbf{y} \mid z \sim \mathcal{N}(z\mu_y, \sigma_y^2)$. Bayes' rule gives

$$\mathrm{PMI}(\mathbf{x},\mathbf{y}) = \log[1 + u(\mathbf{x})v(\mathbf{y})], \qquad u(\mathbf{x}) = \tanh(\mu_x\mathbf{x}/\sigma_x^2), \quad v(\mathbf{y}) = \tanh(\mu_y\mathbf{y}/\sigma_y^2).$$

For nonzero means, the expansion $\mathrm{PMI}(\mathbf{x},\mathbf{y}) = \sum_{j\ge1}(-1)^{j+1}u(\mathbf{x})^j v(\mathbf{y})^j/j$ has infinite rank.

We use $\mu_x = \mu_y = 1.2$ and $\sigma_x = \sigma_y = 1$. Gauss–Hermite quadrature under the mixture marginals gives the following approximate eigenvalues of T^*T:

j	1	2	4	8	12	20
λ_j	1.291	4.763×10^{-1}	4.056×10^{-3}	1.592×10^{-5}	2.257×10^{-7}	1.973×10^{-10}

The observed spectrum decays rapidly despite infinite rank (of order $10^{-i/2}$ approximately)..

Results. we evaluate (i) the rank-D spectral truncation for Theorem 4.3 and (ii) two ReLU encoders trained with the contrastive objective. We report both L_2 norm and approximation error for contrastive loss.

D	Spectral tail	spectral truncation approximation error	MLP L_2 distance	MLP approximation error
1	4.902×10^{-1}	3.808×10^{-2}	3.763×10^{-1}	9.941×10^{-3}
2	1.397×10^{-2}	9.984×10^{-4}	4.607×10^{-2}	4.616×10^{-4}
4	6.822×10^{-4}	5.638×10^{-5}	3.952×10^{-3}	1.877×10^{-5}
8	6.165×10^{-6}	5.587×10^{-7}	1.529×10^{-3}	9.188×10^{-6}

Table 2: Approximation errors for infinite-rank Gaussian mixture model.

As we can see, the approximation error gets smaller as D grows for both spectral truncation and MLP.

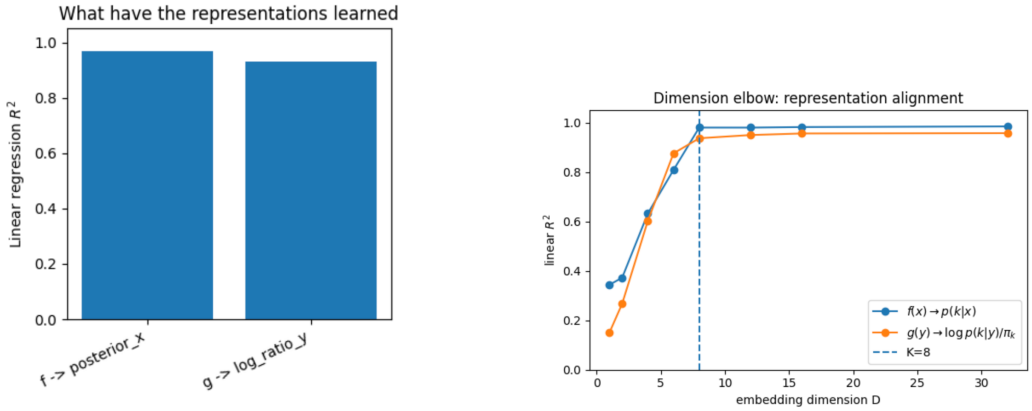
I.2 Neural-network experiments under LCGM

We next examine the structured-data predictions in Section 5.2 using an imbalanced Gaussian mixture with a cleaner anchor view. We evaluate whether the two encoders recover the posterior features $p(z = k \mid \mathbf{x})$ and log-posterior-ratio features $\log[p(z = k \mid \mathbf{y})/\pi_k]$, and whether representation quality saturates around the latent-class dimension K.

Setup. A latent class $z \in [K]$ is sampled from a prior distribution π, where $\pi_k \propto 1/(k+1)$. Conditional on $z = k$, the two views are generated independently as

$$\mathbf{x} \mid z = k \sim \mathcal{N}(\mu_k^1, \sigma_1^2 I), \qquad \mathbf{y} \mid z = k \sim \mathcal{N}(\mu_k^2, \sigma_2^2 I).$$

The class centers $\{\mu_k^1\}_{k=1}^K$ and $\{\mu_k^2\}_{k=1}^K$ are sampled independently. We choose $\sigma_1 = 0.35 < \sigma_2 = 1$, so that the $\mathcal{X}$-view is substantially cleaner than the $\mathcal{Y}$-view. As a result, $p(z|\mathbf{x})$ is close to a one-hot latent identity feature, approximating the deterministic setting in (14). We use $K = 8$ latent classes and $d_1 = d_2 = 16$. We train three-layer MLP encoders with hidden dimension 128 using the inner-product score with $\tau = 0.5$. The model is optimized by AdamW for 20 epochs with batch size 512. The experiments are performed on Google colab with T4 GPU. Our code is available at `https://github.com/YuanfanLi2233/approximation_CRL_LCGM_NIPS2026`.



(a) Recovered singular functions.

(b) Dimension elbow.

Figure 1: **Synthetic validation of the representation learning in LCGM.** (a) The two encoders recover the probabilistic features predicted by Proposition 5.6. (b) Representation alignment improves as the embedding dimension increases and saturates around the latent rank $K = 8$.

Evaluation. According to Proposition 5.6, the encoder f should recover the posterior feature $p(z = k|\mathbf{x})$, while the encoder g should recover the log-posterior-ratio feature $\log(p(z = k|\mathbf{y})/\pi_k)$. In the Gaussian mixture model, both quantities can be computed in closed form. We therefore evaluate the learned representations by linear probing: we report the linear regression R^2 from $f(\mathbf{x})$ to $p(z = k|\mathbf{x})$ and from $g(\mathbf{y})$ to $\log(p(z = k|\mathbf{y})/\pi_k)$. We use R^2 because Proposition 5.6 identifies the relevant singular-function spaces only up to an invertible linear transformation; hence a high linear-regression R^2 indicates that the learned representation spans the predicted feature space. We also vary the embedding dimension D.

Representations recover singular functions. Figure 1a shows what the two representations learn. The clean-side encoder f linearly predicts the posterior feature $p(z = k|\mathbf{x})$ with high accuracy, achieving $R^2 \approx 0.97$. The noisy-side encoder g linearly predicts the log-posterior-ratio feature $\log(p(z = k|\mathbf{y})/\pi_k)$ with $R^2 \approx 0.93$. This supports the spectral interpretation in Proposition 5.6: the learned representations recover singular functions of the PMI operator.

Representation quality for different dimensions. Figure 1b further examines the role of the embedding dimension. Under LCGM, the PMI operator has rank at most K, so a K-dimensional inner-product representation is sufficient at the population level. Consistent with this prediction, the representation quality improves as D increases and saturates around $D = K$. When $D < K$, the representation is rank-limited and cannot recover all latent semantic directions.

I.3 Verification of bounded density-ratio and Tsybakov noise condition

In this section, we verify the bounded density-ratio condition in Assumption 4.2 and empirically assess the Tsybakov noise condition in Assumption 6.1 using the Gaussian mixture model in Section I.1.

Bounded density ratio. In the above model, the density ratio

$$\frac{p(\mathbf{y} \mid \mathbf{x})}{p_{\mathcal{Y}}(\mathbf{y})} = 1 + \tanh(1.2\mathbf{x}) \tanh(1.2\mathbf{y})$$

lies in $(0, 2)$. Thus Assumption 4.2 holds with $\alpha = 2$.

Tsybakov noise condition. For independent $(\mathbf{x}, \mathbf{y}, \mathbf{y}') \sim P_{\mathcal{X}} \otimes P_{\mathcal{Y}} \otimes P_{\mathcal{Y}}$, we estimate the margin probabilities $\mathbb{P}\Big(\Big|\frac{p(\mathbf{y}|\mathbf{x})}{p_{\mathcal{Y}}(\mathbf{y})} - \frac{p(\mathbf{y}'|\mathbf{x})}{p_{\mathcal{Y}}(\mathbf{y}')}\Big| \le t\Big)$ for different values of t in Assumption 6.1 through Monte Carlo sampling. The estimated probabilities are shown in the following table.

t	$\widehat{\mathbb{P}}\Big(\Big\|\frac{p(\mathbf{y}\|\mathbf{x})}{p_{\mathcal{Y}}(\mathbf{y})} - \frac{p(\mathbf{y}'\|\mathbf{x})}{p_{\mathcal{Y}}(\mathbf{y}')}\Big\| \le t\Big)$
10^{-5}	1.7360×10^{-4}
3×10^{-5}	4.5680×10^{-4}
10^{-4}	1.3220×10^{-3}
3×10^{-4}	3.4742×10^{-3}
10^{-3}	9.7374×10^{-3}

A log–log fit gives the noise exponent $\widehat{\mu} = 0.878890$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clarify our contributions and the scope of the paper in the abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discusses the limitations of the work. For example, in the conclusion, we mention that our analysis can not cover the cosine similarity scoring function.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We present all the necessary assumptions and include the proof of all theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: As a theoretical paper, our main contributions are theoretical analyses and proofs. The numerical experiments are designed solely to provide basic validation of our theoretical findings. We have fully disclosed all experimental details in Section I of the paper.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   The details and code of experiments can be found in Appendix I.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The details of experiments can be found in Section I.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: Our paper is primarily theoretical in nature. The simple numerical experiments are designed as deterministic verifications of our theoretical results rather than empirical evaluations requiring statistical analysis.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Given the simplicity of our numerical experiments (which are minimal and for verification only), we conducted our experiments on Google Colab with T4 GPU.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper confirm with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is a theoretical paper, and there is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This is a theoretical paper, which poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: This is a theoretical paper, and we do not use any existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not release any new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This is a theoretical paper, which does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This is a theoretical paper, which does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.