STATISTICAL GUARANTEES FOR MULTIMODAL CONTRASTIVE LEARNING VIA POINTWISE MUTUAL INFORMATION ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal contrastive learning has emerged as a powerful approach for learning representations from paired multimodal data. In this work, we develop a theoretical framework for analyzing the finite-sample behavior of contrastive learning with the InfoNCE objective. Our approach is based on viewing the learned similarity score as an estimator of the pointwise mutual information (PMI) function, which characterizes the population-optimal similarity score up to an additive constant. By exploiting the curvature of the InfoNCE objective together with a localized analysis around the PMI score, we derive fast convergence rates for the PMI estimation error despite the dependence among negative pairs in InfoNCE. When PMI is Hölder smooth, we further show that InfoNCE achieves the minimax-optimal rate up to logarithmic factors for both joint-input neural networks and CLIP-style dual encoders. For downstream learning, we show that accurate recovery of PMI implies approximate sufficiency of the learned representations and yields a bound on the excess risk of zero-shot classification. We provide numerical experiments to support our theoretical results.

1 INTRODUCTION

Multimodal contrastive learning has become a widely used approach for learning representations across modalities and has enabled significant advances in applications including zero-shot visual recognition (Zhai et al., 2022; Mu et al., 2022), text-guided image generation (Ramesh et al., 2022; Crowson et al., 2022), and cross-modal retrieval (Jia et al., 2021; Yao et al., 2022). Building on contrastive objectives developed for representation learning (Oord et al., 2018; Chen et al., 2020), methods such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) learn a pair of encoders that assign high similarity to positive pairs and low similarity to negative pairs. The resulting representations can then be transferred to a wide range of downstream tasks without training new modality-specific encoders from scratch (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2022). Despite their widespread empirical success, however, the theoretical understanding of multimodal contrastive learning remains limited, particularly regarding its finite-sample statistical behavior.

A useful way to understand multimodal contrastive learning is to characterize the similarity score learned by the encoders at the population level. For the widely used InfoNCE objective (Oord et al., 2018), the population-optimal similarity score is given by the pointwise mutual information (PMI), up to an additive constant, which measures the dependence between the two modalities (Uesaka et al., 2025; Oko et al., 2025). This characterization provides a statistical interpretation of contrastive pretraining as directly estimating the dependence between modalities through the similarity score. However, much less is known about how accurately contrastive learning estimates this PMI function from finitely many paired observations, and how the accuracy of PMI estimation translates into the quality of the learned representations and their downstream performance. In particular, because PMI determines the conditional distribution of one modality given the other, it provides a natural population quantity underlying many downstream uses of contrastive representations.

In this work, we develop a theoretical framework for multimodal contrastive learning by viewing InfoNCE as a finite-sample estimator of PMI. This perspective allows us to move beyond generalization of the contrastive loss and directly quantify recovery of the dependence structure between the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

two modalities. Our analysis is based on two key components. First, we develop a self-concordance-type argument for the softmax loss, analogous to that for logistic regression in Bach (2010), which controls the curvature of the InfoNCE loss and relates excess InfoNCE risk to the estimation error of PMI. Second, although the same observations are repeatedly used as negatives within a batch, inducing within-batch dependence, we show that the resulting interactions are sufficiently weak to apply concentration inequalities for functions of bounded interaction (Maurer, 2019). Combining this concentration result with a localized analysis, we obtain fast finite-sample bounds governed by the total number of observed pairs rather than the number of batches. Specializing these bounds to Hölder-smooth PMI yields convergence rates that are minimax optimal up to logarithmic factors. We further derive corresponding downstream guarantees for the quality of the learned representations and zero-shot classification. Our main contributions are summarized as follows.

- We establish a general finite-sample oracle inequality for InfoNCE that directly controls the estimation error of PMI and yields fast rates through a localized complexity analysis.
- We specialize the general theory to both joint-input neural networks and CLIP-style dual encoders. For Hölder-smooth PMI, the resulting estimators achieve the minimax rate up to logarithmic factors, and we derive a matching minimax lower bound.
- For downstream learning, we show that the estimation error of PMI provides theoretical guarantees for the approximate sufficiency of the learned representations and for the excess risk of zero-shot classification.

2 Related work

Multimodal contrastive learning. CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) established dual-encoder contrastive pretraining as a powerful approach for learning transferable image-text representations from large-scale paired data. Subsequent work has extended this framework by modifying the contrastive objective, the cross-modal similarity, or the representation space. For example, CLOOB (Fürst et al., 2022) combines Hopfield networks with the InfoLOOB objective, FILIP (Yao et al., 2022) introduces token-wise late interaction for fine-grained image-text alignment, and SigLIP (Zhai et al., 2023) replaces the softmax contrastive objective with pairwise sigmoid losses. Other approaches go beyond standard Euclidean representations through hyperbolic geometry (Desai et al., 2023) or adopt more flexible similarity classes such as weighted point set embeddings (Uesaka et al., 2025). Beyond these methodological developments, theoretical work has characterized population contrastive objectives from several perspectives. For example, Wang & Isola (2020) describe contrastive representations through alignment and uniformity, Zimmermann et al. (2021); Daunhawer et al. (2023) establish identifiability results under latent generative models, and Gui et al. (2025) show that multimodal contrastive learning can adapt to the intrinsic dimension of shared latent variables.

Statistical theory of contrastive learning. There has been a growing body of work on the finite-sample theory of contrastive learning. Early work relates contrastive pretraining to downstream classification under structural assumptions (Saunshi et al., 2019), with subsequent analyses studying the role of negative samples and the gap between contrastive and supervised losses (Ash et al., 2022; Bao et al., 2022). Related guarantees for downstream prediction have been developed through multi-view redundancy and augmentation graphs (Tosh et al., 2021; HaoChen et al., 2021). More recent work develops finite-sample guarantees in a range of settings. For example, Nakada et al. (2023) study feature recovery for multimodal contrastive learning under linear representation models, while Chen et al. (2024) provide finite-sample guarantees for CLIP and derive zero-shot transfer guarantees under a shared-feature model. Lei et al. (2023) obtain generalization bounds with only logarithmic dependence on the number of negatives, and faster rates when the empirical risk is near zero, while subsequent work improves the dependence on network depth (Hieu et al., 2025) and handles the dependence among negative pairs (Hieu & Ledent, 2025; van Elst & Ghoshdastidar, 2025). Li et al. (2026) show that the contrastive loss is statistically consistent with the optimal ranking under an AUC-type retrieval criterion, and establish generalization bounds for contrastive learning. More closely related to our setting, Oko et al. (2025) show that near-minimizers of the CLIP objective yield approximate sufficient representations and derive sample-complexity bounds for multimodal downstream tasks under a generative hierarchical model. Building on this framework, Lin & Mei

(2025) extend the notion of approximate sufficiency and obtain finite-sample guarantees for data-augmentation-based contrastive learning.

PMI and density ratio estimation. Since PMI is the log-density ratio between the joint distribution and the product of its marginals, its estimation is closely related to density ratio estimation (Sugiyama et al., 2012; Nguyen et al., 2010; Gutmann & Hyvärinen, 2010). In the framework of density ratio estimation, Tsai et al. (2020) propose neural methods for estimating the joint-to-product density ratio, whose logarithm gives PMI. More recently, Han et al. (2025) estimate PMI using neural Bregman objectives and establish nonasymptotic and minimax rate guarantees. Zheng et al. (2026) develop related finite-sample theory for two sample density ratio estimation with deep neural networks, which can also cover PMI by comparing the joint distribution with the product of marginals. These works directly estimate density ratios or their logarithms, whereas we study the score learned by the contrastive objective, whose population optimum is PMI.

3 Problem setup

3.1 InfoNCE loss and empirical minimizer

We begin by introducing the InfoNCE objective and its empirical minimizer. Let (X, Y) denote a paired observation from two modalities, such as image and text, drawn from a joint distribution P on $\mathcal{X} \times \mathcal{Y}$. We denote the marginal distributions of X and Y by P_X and P_Y, and the product distribution of the marginals by $Q = P_X \otimes P_Y$. In multimodal contrastive learning, a pair (X, Y) drawn from P is used as a positive pair, while a pair (X, Y) drawn from Q is used as a negative pair.

The goal of multimodal contrastive learning is then to learn a similarity score $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ that assigns high scores to positive pairs and low scores to negative pairs. For example, in CLIP (Radford et al., 2021), the score is parameterized by two encoders as $s(x, y) = \langle f(x), g(y) \rangle / \tau$ with $f : \mathcal{X} \to \mathbb{R}^p$, $g : \mathcal{Y} \to \mathbb{R}^p$, and a temperature $\tau > 0$. Given a batch $B = ((X_i, Y_i))_{i=1}^K$ of K independent pairs from P, the score can be learned by minimizing the InfoNCE loss (Oord et al., 2018)

$$\ell_K(s; B) = -\frac{1}{2K} \sum_{i=1}^{K} \log \frac{e^{s(X_i, Y_i)}}{K^{-1} \sum_{j=1}^{K} e^{s(X_i, Y_j)}} - \frac{1}{2K} \sum_{i=1}^{K} \log \frac{e^{s(X_i, Y_i)}}{K^{-1} \sum_{j=1}^{K} e^{s(X_j, Y_i)}}. \tag{1}$$

Note that in (1), negative pairs (X_i, Y_j), $i \neq j$, appear once when X_i is compared with the Y's in the batch and once when Y_j is compared with the X's. The corresponding population loss is

$$L_K(s) = \mathbb{E}[\ell_K(s; B)], \tag{2}$$

where the expectation is over K independent pairs in B.

In practice, we observe a finite sample $(X_i, Y_i)_{i=1}^n$ of $n = NK$ independent pairs from P, which are divided into N disjoint batches $B_1, \ldots, B_N$ of size K. Given a class of score functions $\mathcal{S}$, we define the empirical InfoNCE loss and its minimizer by

$$\widehat{L}_K(s) := \frac{1}{N} \sum_{b=1}^{N} \ell_K(s; B_b), \qquad \widehat{s} = \arg\min_{s \in \mathcal{S}} \widehat{L}_K(s). \tag{3}$$

The resulting score $\widehat{s}(x, y)$ can be used to compare new pairs from the two modalities, for example in cross-modal retrieval by ranking candidate texts for a given image. For CLIP-style parameterizations, minimizing the InfoNCE loss also learns the encoders f and g, whose representations can be used for downstream tasks such as zero-shot classification (Radford et al., 2021).

3.2 Pointwise mutual information and population minimizers

Next, we characterize the optimal similarity score for the population InfoNCE loss. Assume that P and Q are mutually absolutely continuous, and define the logarithmic density ratio

$$s^\star(x, y) := \log \frac{dP}{dQ}(x, y). \tag{4}$$

This function compares the likelihood of observing (x, y) under the joint distribution with that under independent pairing, and is also known as the pointwise mutual information (PMI). The following proposition shows that PMI, up to an additive constant, minimizes the population InfoNCE loss (Oord et al., 2018; Oko et al., 2025; Lin & Mei, 2025).

Proposition 3.1. *For every $K \geq 2$, the minimizers of the population InfoNCE loss L_K (2) over measurable scores are given by $\{s^\star + c : c \in \mathbb{R}\}$, Q-almost surely.*

The proof for discrete $\mathcal{X}$ and $\mathcal{Y}$ is given in Oko et al. (2025, Lemma 1). The same argument extends to the continuous setting, and we provide the proof in Appendix A.2 for completeness.

According to Proposition 3.1, the empirical InfoNCE minimizer $\widehat{s}$ can be viewed as an estimator of $s^\star$ up to an additive constant. Since $\ell_K(s + c; B) = \ell_K(s; B)$ for every constant c, this constant cannot be identified from the InfoNCE loss. We therefore measure the error of a score s relative to $s^\star$ after removing this ambiguity, i.e.,

$$\mathrm{Err}_Q(s) = \inf_{c \in \mathbb{R}} \|s - s^\star - c\|^2_{L^2(Q)} = \mathrm{Var}_Q(s - s^\star). \tag{5}$$

Note that while InfoNCE identifies the score only up to an additive constant, this ambiguity does not affect the downstream tasks considered in Section 5.

To facilitate the finite-sample analysis, we assume throughout that all score functions in the feasible score function class $\mathcal{S}$, as well as the PMI function, are bounded.

Assumption 3.2 (Bounded scores)**.** *There exists a constant $M_0 \geq 1$ such that*

$$\sup_{s \in \mathcal{S}} \|s\|_\infty \leq M_0, \qquad \|s^\star\|_\infty \leq M_0.$$

Similar boundedness assumptions are commonly used in finite-sample analyses of contrastive learning; see, e.g., Oko et al. (2025); Lin & Mei (2025); Li et al. (2026). The constants in our finite-sample bounds may depend on M_0, but not on the sample size n or the batch size K.

4 Finite-sample guarantees for contrastive learning

In this section, we derive upper bounds on the finite-sample error of the empirical InfoNCE minimizer. We first establish a general oracle bound for any function class and then specialize it to classes of neural networks when PMI is Hölder smooth.

4.1 Finite-sample estimation of pointwise mutual information

Our analysis proceeds by relating population excess risk to the estimation error of PMI and then bounding the difference between empirical and population losses over scores with small excess risk.

Let $\mathcal{E}_K(s) = L_K(s) - L_K(s^\star)$ denote the population excess risk of a score function $s \in \mathcal{S}$. Our goal is to control $\mathrm{Err}_Q(\widehat{s})$, where $\widehat{s}$ is the minimizer of the empirical InfoNCE loss (3). We first establish a connection between the population excess risk and the squared PMI estimation error.

Proposition 4.1. *Under Assumption 3.2, there exists a constant $C > 0$ depending only on M_0 such that for every $K \geq 2$ and $s \in \mathcal{S}$,*

$$\mathrm{Err}_Q(s) \leq C\mathcal{E}_K(s).$$

The proof of Proposition 4.1 is given in Appendix A.3. The key observation is that the InfoNCE loss (1) has the same softmax likelihood structure as multinomial logistic regression, which yields a self-concordance type property analogous to Bach (2010, Lemma 1, Proposition 1). Specifically, for $A(t) = L_K\big(s^\star + t(s - s^\star)\big)$, $0 \leq t \leq 1$, a direct calculation under Assumption 3.2 gives $|A'''(t)| \leq 4M_0 A''(t)$. As in Bach (2010), this prevents the curvature from decaying too quickly along the path and allows us to derive Proposition 4.1 by relating the curvature at $s^\star$ to $\mathrm{Err}_Q(s)$.

We now aim to control the population excess risk of $\widehat{s}$. By the optimality of $\widehat{s}$, for every $s \in \mathcal{S}$, we have $\mathcal{E}_K(\widehat{s}) \leq \mathcal{E}_K(s) + \big\{(L_K - \widehat{L}_K)(\widehat{s}) - (L_K - \widehat{L}_K)(s)\big\}$. The first term depends on how well $\mathcal{S}$ approximates the population optimum, while the second reflects the difference between the

empirical and population losses. To control the second term, we use the fact that scores with small excess risk are close to $s^\star$ in $L^2(Q)$. For $\rho > 0$, define $\mathcal{S}(\rho) = \{s \in \mathcal{S} : \mathcal{E}_K(s) \le \rho^2\}$, and let

$$\Delta_n(\rho) = \sup_{s \in \mathcal{S}(\rho)} \left| (\widehat{L}_K - L_K)(s) - (\widehat{L}_K - L_K)(s^\star) \right|. \tag{6}$$

Due to Proposition 4.1, every $s \in \mathcal{S}(\rho)$ lies within $O(\rho)$ of $s^\star$ in $L^2(Q)$, up to an additive constant. As ρ decreases, the set $\mathcal{S}(\rho)$ becomes smaller, and so the deviation $\Delta_n(\rho)$ cannot increase.

Now let $\mathfrak{R}_n(\rho)$ denote a high-probability upper bound on $\Delta_n(\rho)$ that holds for each fixed N, K with $n = NK$ (see (36) in Appendix B.4 for its precise definition). Following the localization approach (Bartlett et al., 2005; Wainwright, 2019), we define the critical radius by

$$\delta_n := \inf \left\{ \rho > 0 : \mathfrak{R}_n(\rho) \le \rho^2/16 \right\}.$$

Appendices B.4 and C.1 provide an explicit bound for $\mathfrak{R}_n(\rho)$ and derive the corresponding critical radius for the function classes considered in Section 4.2, respectively. The following theorem gives the resulting error bound for $\widehat{s}$, whose proof is given in Appendix B.4.

Theorem 4.2. *Suppose Assumption 3.2 holds. For every $\eta \in (0,1)$, $K \ge 2$ and $N \ge 1$ with $n = NK$, with probability at least $1 - \eta$,*

$$\mathcal{E}_K(\widehat{s}) \le \underbrace{\frac{5}{3} \inf_{s \in \mathcal{S}} \mathcal{E}_K(s)}_{\textit{approximation error}} + \underbrace{\frac{\delta_n^2}{6}}_{\textit{estimation error}}, \qquad \mathrm{Err}_Q(\widehat{s}) \le C \left\{ \inf_{s \in \mathcal{S}} \mathcal{E}_K(s) + \delta_n^2 \right\}, \tag{7}$$

where $C > 0$ depends only on M_0 and not on n, N, or K. If $s^\star + c \in \mathcal{S}$ for some $c \in \mathbb{R}$, then

$$\mathcal{E}_K(\widehat{s}) \le \frac{\delta_n^2}{12}, \qquad \mathrm{Err}_Q(\widehat{s}) \le C\delta_n^2. \tag{8}$$

Note that the bound in (7) has two terms. The approximation term vanishes when $s^\star + c \in \mathcal{S}$ for some $c \in \mathbb{R}$, in which case the constant in the excess risk bound can be improved as in (8). The estimation term is due to finitely many observations and is determined by the critical radius.

4.2 ReLU networks and minimax rates for Hölder-smooth PMI

We next apply Theorem 4.2 to two function classes based on ReLU networks: scalar-valued ReLU networks on the joint input (x, y), and CLIP-style scores $s_{f,g}(x, y) = \langle f(x), g(y) \rangle / \tau$, where both encoders f and g are ReLU networks. We then consider the case where the true PMI $s^\star$ is Hölder smooth and show that both classes achieve the minimax rate up to logarithmic factors.

Let $\mathcal{X} = [0,1]^{d_X}$ and $\mathcal{Y} = [0,1]^{d_Y}$, where $d_X, d_Y \ge 1$. For $p \ge 1$, let $\mathcal{R}_p(d, W, L, B, S)$ denote the class of ReLU networks $h : [0,1]^d \to \mathbb{R}^p$ with depth at most L, width at most W, at most S nonzero weights and biases, and all weights and biases bounded in absolute value by B; see Schmidt-Hieber (2020) for a similar network class. For $p = 1$, we write $\mathcal{R}(d, W, L, B, S) = \mathcal{R}_1(d, W, L, B, S)$. We assume $B \ge 1$, $S \ge \max\{W, L\}$, and $W \ge \max\{d, p\}$ (Ou & Bölcskei, 2026, Theorem 5.1).

Joint-input ReLU scores. We first consider a ReLU network on the joint input (x, y). Define $\mathcal{S}_{\mathrm{net}} = \{s \in \mathcal{R}(d_X + d_Y, W, L, B, S) : \|s\|_\infty \le M_0\}$. To measure how well this network class can approximate $s^\star$ (up to an additive constant), let

$$\epsilon_{\mathrm{net}} = \inf_{s \in \mathcal{S}_{\mathrm{net}},\, c \in \mathbb{R}} \|s - s^\star - c\|_\infty.$$

Let $\widehat{s}$ denote the empirical InfoNCE minimizer over $\mathcal{S}_{\mathrm{net}}$. The following result bounds the error of $\widehat{s}$ in terms of ϵ_{net} and the size of the network class; the proof is provided in Appendix C.1.

Corollary 4.3. *Under Assumption 3.2, for every $\eta \in (0,1)$, $K \ge 2$ and $N \ge 1$ with $n = NK$, with probability at least $1 - \eta$,*

$$\mathrm{Err}_Q(\widehat{s}) \le C \left\{ \epsilon_{\mathrm{net}}^2 + \frac{\min\{S, W^2 L\} \log\big(en(W+1)^L B^L\big) + \log(1/\eta)}{n} \right\},$$

where $C > 0$ depends only on M_0.

CLIP-style scores. Next we consider the CLIP-style score $s_{f,g}(x,y) = \langle f(x), g(y)\rangle/\tau$ with $\tau > 0$. Define $\mathcal{S}_{\mathrm{IP}} = \{s_{f,g} : f \in \mathcal{R}_p(d_X, W, L, B, S),\ g \in \mathcal{R}_p(d_Y, W, L, B, S),\ \|s_{f,g}\|_\infty \le M_0\}$, that is, both encoders are p-dimensional ReLU networks, $f : \mathcal{X} \to \mathbb{R}^p$ and $g : \mathcal{Y} \to \mathbb{R}^p$. To measure how well this class can approximate $s^\star$ (up to an additive constant), let

$$\epsilon_{\mathrm{IP}} = \inf_{s \in \mathcal{S}_{\mathrm{IP}},\, c \in \mathbb{R}} \|s - s^\star - c\|_\infty.$$

Let $\widehat{s}$ denote the empirical InfoNCE minimizer over $\mathcal{S}_{\mathrm{IP}}$. The following result bounds the error of $\widehat{s}$ in terms of ϵ_{IP} and the size of the network class; the proof is provided in Appendix C.1.

Corollary 4.4. *Under Assumption 3.2, for every $\eta \in (0,1)$, $K \ge 2$ and $N \ge 1$ with $n = NK$, with probability at least $1-\eta$,*

$$\mathrm{Err}_Q(\widehat{s}) \le C \left\{ \epsilon_{\mathrm{IP}}^2 + \frac{\min\{S, W^2 L\} \log\left[en \left(1 + \frac{p(W+1)^{2L} B^{2L}}{\tau}\right)\right] + \log(1/\eta)}{n} \right\},$$

where $C > 0$ depends only on M_0.

Hölder-smooth PMI. For $\alpha > 0$, let $r = \lceil \alpha \rceil - 1$ and $\beta = \alpha - r \in (0,1]$. For a function $h : [0,1]^{d_X + d_Y} \to \mathbb{R}$, define the Hölder norm

$$\|h\|_{C^\alpha} := \sum_{|\nu| \le r} \|\partial^\nu h\|_\infty + \sum_{|\nu| = r} \sup_{z \ne z'} \frac{|\partial^\nu h(z) - \partial^\nu h(z')|}{\|z - z'\|_\infty^\beta},$$

where ν is a multi-index and $|\nu|$ denotes its order. We say that h belongs to the Hölder class with smoothness α and radius M if $\|h\|_{C^\alpha} \le M$.

Now consider the case where, for some $\alpha, M > 0$, the true PMI $s^\star$ belongs to this Hölder class, i.e.,

$$\|s^\star\|_{C^\alpha} \le M, \tag{9}$$

and take $M_0 = M + 1$. The following result shows that both the joint-input and CLIP-style ReLU classes achieve the same convergence rate for estimating $s^\star$.

Theorem 4.5. *Suppose $s^\star$ satisfies (9), i.e., it is Hölder smooth with smoothness α and radius M. For all sufficiently large n, the network sizes can be chosen such that for either the joint-input or the CLIP-style score class above, the corresponding empirical InfoNCE minimizer $\widehat{s}$ satisfies, for every $\eta \in (0,1)$, $K \ge 2$ and $N \ge 1$ with $n = NK$, with probability at least $1-\eta$,*

$$\mathrm{Err}_Q(\widehat{s}) \le C \left(\frac{\log^3(en)}{n} \right)^{\frac{2\alpha}{2\alpha + d_X + d_Y}} + C \frac{\log(1/\eta)}{n}.$$

The constant $C > 0$ may depend on α, d_X, d_Y, M and, for the CLIP-style class, on the fixed temperature τ, but not on n, N, K.

The proof is given in Appendix C.2. The result follows by combining Corollaries 4.3 and 4.4 with sparse ReLU approximation of Hölder functions for the joint-input class and a corresponding two-encoder approximation for the CLIP-style class.

Minimax optimality. The following lower bound shows that the exponent of n in Theorem 4.5 cannot be improved when estimating PMI from paired observations.

Theorem 4.6. *Fix $\alpha > 0$ and $M > 0$. There exists a constant $c > 0$, depending only on α, d_X, d_Y, and M, such that for every $n \ge 1$,*

$$\inf_{\widetilde{s}} \sup_{P :\, \|s^\star\|_{C^\alpha} \le M} \mathbb{E}_{P^{\otimes n}} \left[\mathrm{Err}_Q(\widetilde{s}) \right] \ge c\, n^{-\frac{2\alpha}{2\alpha + d_X + d_Y}},$$

where, for each joint distribution P, $Q = P_X \otimes P_Y$ and $s^\star = \log(dP/dQ)$. The infimum is over all measurable estimators based on n independent paired observations from P.

The proof is given in Appendix C.3. Theorem 4.5 and Theorem 4.6 show that the empirical InfoNCE estimators over both the joint-input and CLIP-style ReLU classes achieve the minimax rate up to logarithmic factors, and no estimator can substantially improve the dependence on n. In the lower-bound construction for Theorem 4.6, both marginals P_X and P_Y are fixed to known uniform distributions. This shows that the difficulty of estimating PMI comes from estimating the dependence between the two modalities rather than the marginal distributions.

4.3 Comparison with related work: global versus local analyses

Recently, Oko et al. (2025) establish finite-sample guarantees for approximate sufficiency in multimodal contrastive learning based on the InfoNCE objective. Under a joint generative hierarchical model, they derive an excess risk bound for CLIP, with their Remark 6 further improving the generalization bound to depend on the total sample size $n = NK$ rather than the number of batches. Relatedly, Lin & Mei (2025) study contrastive learning based on data augmentation with the SimCLR objective and derive a finite-sample bound whose generalization term also depends on n. For a fixed function class $\mathcal{S}$, the generalization bounds in these works are global, i.e., the difference between empirical and population losses is controlled uniformly over $\mathcal{S}$ at a rate of order $\widetilde{O}(\sqrt{\mathrm{Complexity}(\mathcal{S})/n})$. In the well-specified setting, i.e., $s^\star + c \in \mathcal{S}$ for some $c \in \mathbb{R}$, combining such a bound with Proposition 4.1 gives an $\widetilde{O}(n^{-1/2})$ rate for the squared PMI error, whereas our Corollaries 4.3 and 4.4 give an $\widetilde{O}(n^{-1})$ rate.

To obtain such a fast rate, we localize the analysis to score functions with small excess risk. A challenge is that the terms in the InfoNCE loss (1) within the same batch are dependent, because each observation is reused multiple times in computing the loss. We show that replacing one paired observation changes the loss difference between two score functions by order $1/K$, while the interaction between two such replacements is of order $1/K^2$. These bounds allow us to apply the Bernstein-type inequality of Maurer (2019) for functions with weak interactions. Combined with localization, this gives (see (6))

$$\Delta_n(\rho) \lesssim \rho\sqrt{\frac{\mathrm{Complexity}(\mathcal{S})}{n}} + \frac{\mathrm{Complexity}(\mathcal{S})}{n},$$

up to logarithmic factors, yielding a $\widetilde{O}(n^{-1})$ rate for the estimation error. In the setting of Hölder-smooth PMI, the complexity of the function class $\mathcal{S}$ grows with n, and balancing this estimation error with the approximation error gives the minimax-optimal rate in Theorem 4.5.

5 Implications for representation learning and zero-shot classification

In this section, we provide guarantees for two downstream consequences of estimating PMI, i.e., approximate sufficiency of the learned representations and zero-shot linear classification. Throughout this section, we focus on CLIP-style scores $s_{f,g}(x, y) = \langle f(x), g(y)\rangle/\tau$ with encoders f and g.

5.1 Approximate sufficiency of learned representations

For an encoder f, approximate sufficiency measures the information loss when $f(X)$ is used instead of X for predicting Y. Following Oko et al. (2025); Lin & Mei (2025), we define $\mathrm{Suff}(f) := I(X;Y) - I(f(X);Y)$. Since $f(X)$ is a function of X, this quantity can equivalently be written as $\mathbb{E}_X D_{\mathrm{KL}}(P_{Y|X} \| P_{Y|f(X)})$. In particular, $\mathrm{Suff}(f) = 0$ means that $P_{Y|X} = P_{Y|f(X)}$ almost surely, or equivalently, $Y \perp\!\!\!\perp X \mid f(X)$, which coincides with the usual notion of sufficiency in classical statistics (Keener, 2010). Approximate sufficiency has been used to establish guarantees for a broad range of downstream tasks (Oko et al., 2025; Lin & Mei, 2025).

Note that PMI determines the conditional distribution through

$$P_{Y|X=x}(dy) = \frac{e^{s^\star(x,y)} P_Y(dy)}{\int e^{s^\star(x,y')} P_Y(dy')}. \tag{10}$$

Replacing $s^\star$ in (10) by the learned score $\widehat{s}(x, y) = \langle \widehat{f}(x), \widehat{g}(y)\rangle/\tau$ then defines a conditional distribution whose dependence on x is only through $\widehat{f}(x)$. Therefore, if $\widehat{s}$ accurately estimates $s^\star$, little information about Y should be lost when $\widehat{f}(X)$ is used in place of X. In particular, $\mathrm{Suff}(\widehat{f})$ should be small when the error of $\widehat{s}$ in estimating PMI is small. The following result makes this precise.

Proposition 5.1. *Let $\widehat{s}(x, y) = \langle \widehat{f}(x), \widehat{g}(y)\rangle/\tau$ be the learned score. Under the conditions of Corollary 4.4, there exists a constant $C > 0$ depending only on M_0 such that*

$$\mathrm{Suff}(\widehat{f}) \le C\, \mathrm{Err}_Q(\widehat{s}).$$

The same result holds with the roles of X and Y, and the roles of $\widehat{f}$ and $\widehat{g}$, reversed.

The proof is given in Appendix D.1. Proposition 5.1 shows that the quality of the learned representation, as measured by approximate sufficiency, is directly controlled by the accuracy of PMI estimation. Combining this with Corollary 4.4 gives a finite-sample guarantee for the approximate sufficiency of $\widehat{f}$.

5.2 Zero-shot linear classification

We next consider zero-shot classification based on image-text contrastive learning, where X and Y, respectively, represent an image and a text. Let $C \in \{1, \dots, m\}$ be a downstream class label for X. For each class c, let ν_c denote a distribution over text prompts representing that class. For example, if c corresponds to "dog", ν_c may be the empirical distribution over prompts such as "a photo of a dog" and "a dog in the park." Given a score s, we define the class logit as $z_c(x; s) = b_c + \mathbb{E}_{Y \sim \nu_c}[s(x, Y)]$, $c = 1, \dots, m$, where b_c is a fixed class bias. For the learned CLIP-style score $\widehat{s}(x, y) = \langle \widehat{f}(x), \widehat{g}(y) \rangle / \tau$, the logit becomes $z_c(x; \widehat{s}) = \langle \widehat{f}(x), \mathbb{E}_{\nu_c}[\widehat{g}(Y)]/\tau \rangle + b_c$, so the classifier is linear in the image representation $\widehat{f}(x)$. This setting covers common choices of prompts for zero-shot classification—for example, the single prompt setting in CLIP corresponds to taking ν_c to be a point mass, while Uesaka et al. (2025) consider class-specific prompt distributions obtained by conditioning the text distribution on disjoint subsets of the text space.

Note that for the true PMI $s^\star$, the logit $z_c(x; s^\star)$ averages the PMI over prompts representing class c. Under suitable conditions on the class-specific prompt distributions, this logit equals $\log P(C = c \mid X = x)$ up to a class-independent term; see Appendix D.2 for a sufficient condition and Uesaka et al. (2025) for related conditions. To quantify the performance of the classifier, for logits $z = (z_1, \dots, z_m)$, let $R(z) = \mathbb{E}[-z_C(X) + \log \sum_{j=1}^m e^{z_j(X)}]$ denote the population cross-entropy loss and let $R^\star = \inf_z R(z) = \mathbb{E}[-\log P(C \mid X)]$ be the optimal population risk. The excess risk can then be written as

$$R(z) - R^\star = \mathbb{E}_X D_{\mathrm{KL}}\big(P(C = \cdot \mid X) \| \ \mathrm{softmax}(z(X))\big),$$

where $\mathrm{softmax}(z)_c = e^{z_c} / \sum_{j=1}^m e^{z_j}$. Thus, the excess risk measures how far the class probabilities based on the logits are from the true downstream class probabilities. With this setup, we obtain the following excess risk guarantee for downstream classification.

Proposition 5.2. *Let $\widehat{s}(x, y) = \langle \widehat{f}(x), \widehat{g}(y) \rangle / \tau$, and suppose that the constant $\widetilde{C}$ defined in Appendix D.2 is finite. Then*

$$R(z(\cdot; \widehat{s})) - R^\star \le 2\{R(z(\cdot; s^\star)) - R^\star\} + \frac{3}{4} \widetilde{C} \, \mathrm{Err}_Q(\widehat{s}).$$

The proof is given in Appendix D.2. Here $\widetilde{C}$ measures how well the class-specific prompt distributions ν_c are covered by the pretraining text distribution P_Y; see (48) in Appendix D.2 for its precise definition and examples. The first term on the right is the classification error that remains even when the true PMI $s^\star$ is used, while the second term is due to the error in estimating PMI.

6 Numerical experiments

We conduct numerical experiments on synthetic data to empirically illustrate the finite-sample behavior predicted by our theoretical results. For each (n, K), we use n independent paired observations and divide them into $N = n/K$ disjoint batches of size K, as in our theoretical setting. We minimize the empirical InfoNCE loss (3) over these fixed batches, while keeping the network architecture and optimization settings fixed across (n, K) to isolate the statistical effects of n and K. Additional experiments on real data are provided in Appendix E.

Experimental setting. We consider both a joint-input ReLU score and a normalized CLIP-style score, using a separate synthetic data generating model for each experiment. For the joint-input score, we generate $X \sim \mathrm{Unif}([0, 1]^5)$ and set $Y = (X + Z) \bmod 1$, where Z is independent of X and has independent coordinates with density $q_\lambda(z) \propto \exp\{\lambda \cos(2\pi z)\}$ on the unit circle. Both

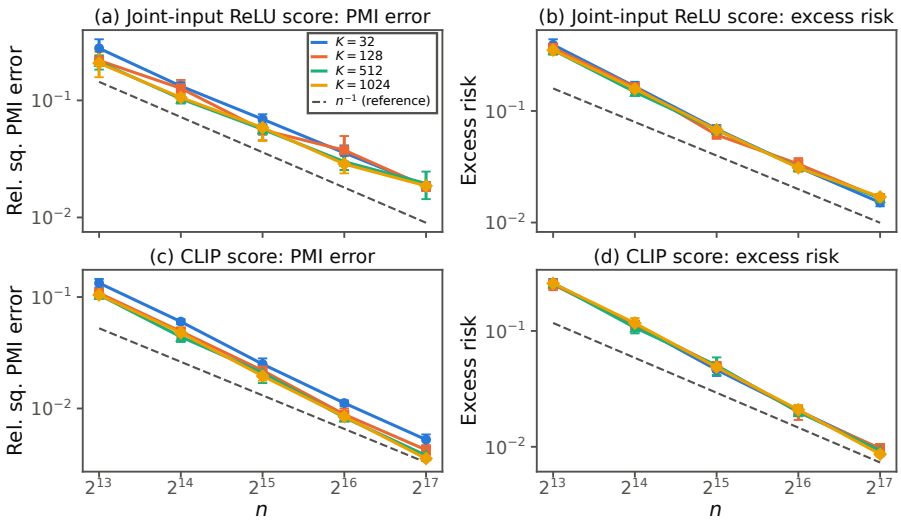

Figure 1: Relative squared PMI error (left) and excess risk (right) versus the total sample size n for different batch sizes K. The top row shows results for the joint-input ReLU score, while the bottom row shows results for the normalized CLIP-style score. Curves show the mean with standard deviation over 5 independent trials. Dashed lines show the reference scaling of n^{-1}.

X and Y are then uniform on $[0,1]^5$, and the resulting PMI is $s^\star(x,y) = \lambda \sum_{j=1}^{5} \cos\{2\pi(x_j - y_j)\} + \text{const}$. We parameterize the score as $s_\theta(x,y) = M_0 \tanh\{h_\theta(x,y)/M_0\}$, where h_θ is a ReLU network with two hidden layers of width 64 and a scalar output, and M_0 is fixed. For the CLIP-style score, we generate $U \sim \text{Unif}(\mathbb{S}^9)$ and $V \mid U = u \sim \text{vMF}(u,\omega)$, where $\text{vMF}(u,\omega)$ is the von Mises-Fisher distribution, and set $X = h_X(U)$ and $Y = h_Y(V)$ using fixed invertible nonlinear transformations. Since these transformations are invertible, the PMI is preserved and is given by $s^\star(x,y) = \omega u^\top v + \text{const}$, where $u = h_X^{-1}(x)$ and $v = h_Y^{-1}(y)$. Each of f and g is a ReLU network with two hidden layers of width 32 and a 20-dimensional output, and we use the normalized score $s_{f,g}(x,y) = \langle f(x)/\|f(x)\|_2, g(y)/\|g(y)\|_2 \rangle/\tau$ with fixed temperature $\tau = 1/\omega$.

For both experiments, we vary $n \in \{2^{13}, \ldots, 2^{17}\}$ and $K \in \{32, 128, 512, 1024\}$. We train using full-batch Adam with learning rate 10^{-3} for 15,000 steps for the joint-input score and 50,000 steps for the CLIP-style score.

Results. Figure 1 summarizes the results. In both experiments, the squared PMI error and the excess risk decrease rapidly with the total sample size n, while the curves for different $K \in \{32, 128, 512, 1024\}$ remain close to each other. Therefore, although the number of batches $N = n/K$ varies substantially with K, the statistical performance is mainly determined by the total number of paired observations n, which is consistent with Theorem 4.2. Additionally, since the network architecture is fixed as n varies, the dashed lines show the reference scaling of n^{-1} for the estimation error as predicted by our theory. We observe that the decay is broadly consistent with this scaling for both the joint-input ReLU and normalized CLIP-style scores, further supporting our theoretical results.

7 Discussion

In this paper, we develop a statistical framework for multimodal contrastive learning by viewing InfoNCE as an estimator of pointwise mutual information. Our theoretical results show that the estimation error is governed by the total number of paired observations despite the dependence created within each batch, and yield minimax-optimal rates up to logarithmic factors for Hölder-smooth PMI. We further show that accurate PMI estimation leads to representations that are approximately sufficient and provides guarantees for zero-shot classification. An interesting direction for future work is to extend the analysis to practical training schemes with stochastic optimization and repeatedly reshuffled batches.

AI USE STATEMENT

In this work, we used generative AI tools as an interactive aid in the development of the paper. The authors formulated the main research questions and mathematical results. Generative AI was used to refine mathematical claims and check arguments for possible gaps or errors. The proofs and final theoretical results were developed and verified by the authors, who critically assessed and, where appropriate, modified AI-assisted suggestions. Generative AI was also used during the writing process to help organize, articulate, and polish mathematical arguments. The final presentation and text were revised and approved by the authors.

For the experiments, the authors determined the research questions, experimental goals, and overall experimental directions. Generative AI was used interactively to suggest possible synthetic and real-data experimental designs and variants. The authors evaluated these suggestions, modified them as needed, and selected the experiments reported in the paper. Generative AI also assisted with portions of the code implementation. All AI-assisted code was inspected, tested, and verified by the authors, and the resulting experimental outputs were independently checked and validated by the authors.

We did not use generative AI for translation. The literature review was led by the authors, with generative AI used interactively to help identify additional potentially relevant literature. All cited works were checked by the authors against the original sources. All AI-assisted work was critically reviewed by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

All assumptions for our theoretical results are stated explicitly in the main paper, and complete proofs are provided in Appendices A to D. Appendix E gives additional details for the synthetic and real-data experiments. The supplementary materials contain the source code for all experiments and figures, experiment configurations, random seeds, per-run results, and scripts for regenerating the reported figures. They also include the real-data preparation and preprocessing pipeline and the metadata used for zero-shot evaluation.

REFERENCES

Jordan Ash, Surbhi Goel, Akshay Krishnamurthy, and Dipendra Misra. Investigating the role of negatives in contrastive representation learning. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 7187–7209. PMLR, 2022.

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4: 384–414, 2010. doi: 10.1214/09-EJS521.

Han Bao, Yoshihiro Nagano, and Kento Nozawa. On the surrogate gap between contrastive and supervised losses. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1585–1606. PMLR, 2022.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005. doi: 10.1214/009053605000000282.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. doi: 10.1093/acprof:oso/9780199535255.001.0001.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020.

Zixiang Chen, Yihe Deng, Yuanzhi Li, and Quanquan Gu. Understanding transferable representation learning and zero-shot transfer in CLIP. In *International Conference on Learning Representations*, pp. 55415–55444, 2024.

Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. VQGAN-CLIP: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision*, pp. 88–105, 2022.

Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E. Vogt. Identifiability results for multimodal contrastive learning. In *International Conference on Learning Representations*, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7694–7731. PMLR, 2023.

Sjoerd Dirksen. Tail bounds via generic chaining. *Electronic Journal of Probability*, 20(53):1–29, 2015. doi: 10.1214/EJP.v20-3760.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pp. 178, 2004.

Andreas Fürst, Elisabeth Rumetshofer, Johannes Lehner, Viet T. Tran, Fei Tang, Hubert Ramsauer, David Kreil, Michael Kopp, Günter Klambauer, Angela Bitto, and Sepp Hochreiter. CLOOB: Modern Hopfield networks with InfoLOOB outperform CLIP. In *Advances in Neural Information Processing Systems*, volume 35, pp. 20450–20468, 2022.

Yu Gui, Cong Ma, and Zongming Ma. Multi-modal contrastive learning adapts to intrinsic dimensions of shared latent variables. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 297–304. PMLR, 2010.

Dongxiao Han, Siming Zheng, Guohao Shen, Xinyuan Song, Liuquan Sun, and Jian Huang. Deep mutual density ratio estimation with Bregman divergence and its applications. *Journal of the American Statistical Association*, 120(551):1990–2001, 2025. doi: 10.1080/01621459.2025.2507437.

Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *Advances in Neural Information Processing Systems*, volume 34, pp. 5000–5011, 2021.

Nong Minh Hieu and Antoine Ledent. Generalization analysis for supervised contrastive representation learning under non-IID settings. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 23179–23218. PMLR, 2025.

Nong Minh Hieu, Antoine Ledent, Yunwen Lei, and Cheng Yeaw Ku. Generalization analysis for deep contrastive representation learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(16):17186–17194, 2025.

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 *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 2021.

Robert W. Keener. *Theoretical Statistics: Topics for a Core Course*. Springer Texts in Statistics. Springer, New York, 2010. doi: 10.1007/978-0-387-93839-4.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yunwen Lei, Tianbao Yang, Yiming Ying, and Ding-Xuan Zhou. Generalization analysis for contrastive representation learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19200–19227. PMLR, 2023.

Yuanfan Li, Xiyuan Wei, Tianbao Yang, and Yiming Ying. Statistical consistency and generalization of contrastive representation learning. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026.

Licong Lin and Song Mei. A statistical theory of contrastive learning via approximate sufficient statistics. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Andreas Maurer. A Bernstein-type inequality for functions of bounded interaction. *Bernoulli*, 25 (2):1451–1471, 2019. doi: 10.3150/18-BEJ1026.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. SLIP: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pp. 529–544, 2022.

Ryumei Nakada, Halil Ibrahim Gulluk, Zhun Deng, Wenlong Ji, James Zou, and Linjun Zhang. Understanding multimodal contrastive learning and incorporating unpaired data. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 4348–4380. PMLR, 2023.

XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010. doi: 10.1109/TIT.2010.2068870.

Kazusato Oko, Licong Lin, Yuhang Cai, and Song Mei. A statistical theory of contrastive pre-training and multimodal generative AI. *arXiv preprint arXiv:2501.04641*, 2025.

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

Weigutian Ou and Helmut Bölcskei. Covering numbers for deep ReLU networks with applications to function approximation and nonparametric regression. *Foundations of Computational Mathematics*, 2026. doi: 10.1007/s10208-026-09762-w.

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 *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5628–5637. PMLR, 2019.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020. doi: 10.1214/19-AOS1875.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2443–2449, 2021.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012. doi: 10.1017/CBO9781139035613.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pp. 1179–1206. PMLR, 2021.

Yao-Hung Hubert Tsai, Han Zhao, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Neural methods for point-wise dependency estimation. In *Advances in Neural Information Processing Systems*, volume 33, pp. 62–72, 2020.

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. In *International Conference on Learning Representations*, pp. 32997–33018, 2025.

Anna van Elst and Debarghya Ghoshdastidar. Tight PAC-Bayesian risk certificates for contrastive learning. *SIAM Journal on Mathematics of Data Science*, 7(4):1904–1927, 2025. doi: 10.1137/24M1715283.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2019. doi: 10.1017/9781108627771.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 2020.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*, 2022.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. LiT: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18123–18133, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Siming Zheng, Guohao Shen, Yuanyuan Lin, and Jian Huang. Error analysis for deep ReLU feedforward density-ratio estimation with Bregman divergence. *Journal of Machine Learning Research*, 27(15):1–60, 2026.

Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12979–12990. PMLR, 2021.

A Properties of the population InfoNCE loss

In this section, we prove Propositions 3.1 and 4.1, which establish key properties of the population InfoNCE loss. Before turning to the proofs, we introduce some notation used throughout the population analysis.

Table 1: Frequently used notation throughout the proofs.

Notation	Definition	Defined in
P, P_X, P_Y	joint law of (X, Y) and its marginals	Section 3.1
$Q = P_X \otimes P_Y$	product of the marginals	Section 3.1
K, N, $n = NK$	batch size, number of batches, number of pairs	Section 3.1
B	a batch of K pairs	Section 3.1
$s^\star = \log(dP/dQ)$	pointwise mutual information	(4)
$r^\star = e^{s^\star}$	density ratio dP/dQ	Appendix A
M_0	bound on $\|s\|_\infty$, $s \in \mathcal{S}$, and on $\|s^\star\|_\infty$	Assumption 3.2
$\kappa = 4M_0$	upper bound on $\mathcal{E}_K(s)$, $s \in \mathcal{S}$ (see Appendix B.1)	Appendix A
$r_{\min} = e^{-M_0}$, $r_{\max} = e^{M_0}$	lower and upper bounds on $r^\star$	(24)
π_i^s, $\bar{\pi}_j^s$	softmax distributions along row i and column j of a batch; $\pi_i^\star$, $\bar{\pi}_j^\star$ at $s = s^\star$	(11)
$\ell_K(s; B)$	batch InfoNCE loss	(1)
L_K, $\widehat{L}_K$	population and empirical InfoNCE losses	(2), (3)
$\mathcal{E}_K(s)$	excess risk $L_K(s) - L_K(s^\star)$	Section 4.1
$\mathrm{Err}_Q(s)$	$\mathrm{Var}_Q(s - s^\star)$, the PMI error up to an additive constant	(5)
$h_{s,t}$	$\ell_K(s; \cdot) - \ell_K(t; \cdot)$ for two scores s, t	Appendix B.1
W_s	$N^{-1} \sum_{b=1}^N \{\ell_K(s; B_b) - \ell_K(s^\star; B_b)\} - \mathcal{E}_K(s)$	Appendix B.1
$d_\infty(s,t)$, $d_Q(s,t)$	$2\|s - t\|_\infty$ and $\|s - t - Q(s - t)\|_{L^2(Q)}$ for two scores s, t, written d_∞ and d_Q when s, t are fixed	Appendix B.1
$\mathcal{S}(\rho)$	$\{s \in \mathcal{S} : \mathcal{E}_K(s) \le \rho^2\}$	Section 4.1
$\mathcal{H}(\epsilon)$	$\log \mathcal{N}(\epsilon, \mathcal{S}, \|\cdot\|_\infty)$	Appendix B.3
$\mathfrak{R}_n(\rho)$, δ_n	envelope and critical radius	Section 4.1; explicit form in (36), (37)
$C_1, C_2, \ldots$	constants, fixed once chosen and independent of K, N, n	
C	a constant that may change from line to line	

For any score s, define the row-wise and column-wise softmax distributions

$$\pi_i^s(j) = \frac{e^{s(X_i, Y_j)}}{\sum_{\ell=1}^K e^{s(X_i, Y_\ell)}}, \qquad \bar{\pi}_j^s(i) = \frac{e^{s(X_i, Y_j)}}{\sum_{\ell=1}^K e^{s(X_\ell, Y_j)}}. \tag{11}$$

Thus, π_i^s is the softmax distribution along row i, while $\bar{\pi}_j^s$ is the softmax distribution along column j. At the PMI score $s^\star$, we write

$$r^\star(x, y) = \frac{dP}{dQ}(x, y) = e^{s^\star(x,y)}$$

for the density ratio, and denote the corresponding row and column softmax distributions by

$$\pi_i^\star(j) = \frac{r^\star(X_i, Y_j)}{\sum_{\ell=1}^K r^\star(X_i, Y_\ell)}, \qquad \bar{\pi}_j^\star(i) = \frac{r^\star(X_i, Y_j)}{\sum_{\ell=1}^K r^\star(X_\ell, Y_j)}.$$

Throughout, we set $\kappa = 4M_0$, which provides a uniform upper bound on the population excess risk, i.e., $\mathcal{E}_K(s) \le \kappa$ for every $s \in \mathcal{S}$ (see (28) in Appendix B.1).

A.1 Population curvature

We first study the population loss along line segments through the PMI score. To this end, we begin with derivative identities for the log-sum-exp function, which will then be applied row-wise and column-wise to the InfoNCE loss.

Lemma A.1. *Let $\alpha_1, \dots, \alpha_K > 0$ and $v_1, \dots, v_K \in \mathbb{R}$. Define*

$$f(t) = \log \sum_{j=1}^{K} \alpha_j e^{t v_j}, \qquad p_j(t) = \frac{\alpha_j e^{t v_j}}{\sum_{\ell=1}^{K} \alpha_\ell e^{t v_\ell}},$$

and $\mu(t) = \sum_{j=1}^{K} p_j(t) v_j$. Then

$$f'(t) = \mu(t), \qquad f''(t) = \sum_{j=1}^{K} p_j(t)\{v_j - \mu(t)\}^2, \qquad f'''(t) = \sum_{j=1}^{K} p_j(t)\{v_j - \mu(t)\}^3. \tag{12}$$

In particular if $v_1, \dots, v_K$ lie in an interval of length a, then

$$0 \le f''(t) \le \frac{a^2}{4}, \qquad |f'''(t)| \le a f''(t) \le \frac{a^3}{4}.$$

Proof. Differentiating f gives $f'(t) = \mu(t)$. Differentiating the weights yields $p_j'(t) = p_j(t)\{v_j - \mu(t)\}$, from which the expression for $f''(t)$ follows immediately. Differentiating once more, the terms involving $\mu'(t)$ vanish because $\sum_j p_j(t)\{v_j - \mu(t)\} = 0$, which gives (12).

Since $\mu(t)$ lies in the convex hull of $v_1, \dots, v_K$, we have $|v_j - \mu(t)| \le a$ for every j, and therefore $|f'''(t)| \le a f''(t)$. Finally, $f''(t)$ is the variance of values contained in an interval of length a, so $0 \le f''(t) \le a^2/4$. Hence $|f'''(t)| \le a^3/4$, completing the proof. □

We next apply these identities to each row and column of the InfoNCE loss.

Lemma A.2. *Assume that $\|\Delta\|_\infty < \infty$, set $a = 2\|\Delta\|_\infty$, and define*

$$A(t) = L_K(s^\star + t\Delta).$$

Writing $\Delta_{ij} = \Delta(X_i, Y_j)$, let $\pi_i^t = \pi_i^{s^\star + t\Delta}$ and $\bar\pi_j^t = \bar\pi_j^{s^\star + t\Delta}$. Set $\mu_i^t = \sum_{j=1}^{K} \pi_i^t(j)\Delta_{ij}$ and $\bar\mu_j^t = \sum_{i=1}^{K} \bar\pi_j^t(i)\Delta_{ij}$. Then

$$\begin{aligned}
A'(t) &= \frac{1}{2K}\mathbb{E}\left[\sum_{i=1}^{K}(\mu_i^t - \Delta_{ii}) + \sum_{j=1}^{K}(\bar\mu_j^t - \Delta_{jj})\right], \\
A''(t) &= \frac{1}{2K}\mathbb{E}\left[\sum_{i=1}^{K}\sum_{j=1}^{K}\pi_i^t(j)(\Delta_{ij} - \mu_i^t)^2 + \sum_{j=1}^{K}\sum_{i=1}^{K}\bar\pi_j^t(i)(\Delta_{ij} - \bar\mu_j^t)^2\right], \\
A'''(t) &= \frac{1}{2K}\mathbb{E}\left[\sum_{i=1}^{K}\sum_{j=1}^{K}\pi_i^t(j)(\Delta_{ij} - \mu_i^t)^3 + \sum_{j=1}^{K}\sum_{i=1}^{K}\bar\pi_j^t(i)(\Delta_{ij} - \bar\mu_j^t)^3\right].
\end{aligned} \tag{13}$$

Proof. Applying Lemma A.1 separately to each row and column gives the pointwise derivative formulas. Since the values Δ_{ij} lie in an interval of length a, the derivative of each row or column loss term is bounded in absolute value by a, while its second and third derivatives are bounded by $a^2/4$ and $a^3/4$, respectively, uniformly in t. Dominated convergence therefore justifies differentiation under the expectation and yields the desired results. □

For a probability vector π on $\{1, \dots, K\}$ and a vector $v = (v_1, \dots, v_K)$, let

$$\mathrm{Var}_\pi(v) = \sum_{j=1}^{K} \pi(j)\left(v_j - \sum_{\ell=1}^{K}\pi(\ell)v_\ell\right)^2.$$

For a perturbation Δ, define the finite-batch variance functional

$$\begin{aligned}
V_K(\Delta) = \frac{1}{2K}\mathbb{E}\Big[&\sum_{i=1}^{K}\mathrm{Var}_{\pi_i^\star}\big(\Delta(X_i, Y_1), \dots, \Delta(X_i, Y_K)\big) \\
&+ \sum_{j=1}^{K}\mathrm{Var}_{\bar\pi_j^\star}\big(\Delta(X_1, Y_j), \dots, \Delta(X_K, Y_j)\big)\Big],
\end{aligned}$$

where $\pi_i^\star = \pi_i^{s^\star}$ and $\bar{\pi}_j^\star = \bar{\pi}_j^{s^\star}$ are the row and column softmax distributions at PMI. Thus, $V_K(\Delta)$ averages the softmax-weighted variance of Δ across the rows and columns of a batch.

The next proposition derives a lower bound on the excess risk using a self-concordance-type argument, analogous to Bach (2010, Lemma 1, Proposition 1). Specifically, the third derivative along a score segment is controlled by the second derivative, preventing the curvature from decreasing too quickly away from the PMI score.

Proposition A.3. *For every score $s = s^\star + \Delta$ with $\|\Delta\|_\infty < \infty$,*

$$\mathcal{E}_K(s) \geq \frac{e^{-a} + a - 1}{a^2} V_K(\Delta) \geq \frac{1}{a+2} V_K(\Delta), \qquad a := 2\|\Delta\|_\infty,$$

with the first factor interpreted as $1/2$ when $a = 0$.

Proof. Let $A(t) = L_K(s^\star + t\Delta)$. For each row, all values Δ_{ij} and their weighted mean μ_i^t lie in an interval of length a. Therefore $|\Delta_{ij} - \mu_i^t| \leq a$. Applying this to Lemma A.2, we obtain $|A'''(t)| \leq aA''(t)$. Hence $\frac{d}{dt}\{e^{at}A''(t)\} = e^{at}\{aA''(t) + A'''(t)\} \geq 0$, so $e^{at}A''(t)$ is nondecreasing and $A''(t) \geq e^{-at}A''(0)$ for $0 \leq t \leq 1$. Since $A'(0) = 0$, Taylor's formula with integral remainder gives

$$A(1) - A(0) = \int_0^1 (1-t)A''(t)dt \geq A''(0) \int_0^1 (1-t)e^{-at}dt.$$

One checks that $\int_0^1 (1-t)e^{-at}dt = \frac{e^{-a}+a-1}{a^2}$, with the value $1/2$ at $a = 0$. Since $A''(0) = V_K(\Delta)$, this proves the first inequality. The second inequality follows from the inequality $(a+2)e^{-a} + a - 2 \geq 0$ for $a \geq 0$. □

The same variance calculation also gives an upper bound on the excess risk.

Lemma A.4. *For every bounded score difference Δ,*

$$\mathcal{E}_K(s^\star + \Delta) \leq \frac{1}{2}\|\Delta\|_\infty^2.$$

Proof. Let $A(t) = L_K(s^\star + t\Delta)$ and $a = \|\Delta\|_\infty$. Since $s^\star$ minimizes L_K (Proposition 3.1), A is minimized at $t = 0$, and hence $A'(0) = 0$. By (13), $A''(t)$ is an average of weighted variances of the values Δ_{ij}. Since $|\Delta_{ij}| \leq a$, these values lie in an interval of length at most $2a$, so $0 \leq A''(t) \leq (2a)^2/4 = a^2$ for every t. Therefore, Taylor's formula with integral remainder gives

$$A(1) - A(0) = \int_0^1 (1-t)A''(t)dt \leq a^2 \int_0^1 (1-t)dt = \frac{a^2}{2}.$$

This proves the claim. □

For any $c \in \mathbb{R}$, the invariance of the InfoNCE loss gives $L_K(s) = L_K(s - c)$. Applying Lemma A.4 to $\Delta = s - s^\star - c$ and then taking the infimum over c yields

$$\mathcal{E}_K(s) \leq \frac{1}{2} \inf_{c \in \mathbb{R}} \|s - s^\star - c\|_\infty^2. \tag{14}$$

A.2 Population InfoNCE minimizers

We first define versions of the conditional distributions induced by the density ratio $r^\star$. Set

$$\mathcal{X}_0 = \left\{x : \int r^\star(x,y)P_Y(dy) = 1\right\}, \qquad \mathcal{Y}_0 = \left\{y : \int r^\star(x,y)P_X(dx) = 1\right\}.$$

Since $r^\star Q = P$ has marginals P_X and P_Y, Tonelli's theorem gives $P_X(\mathcal{X}_0) = P_Y(\mathcal{Y}_0) = 1$. We define the conditional kernels

$$P_{Y|X=x}(dy) = \begin{cases} r^\star(x,y)P_Y(dy), & x \in \mathcal{X}_0, \\ P_Y(dy), & x \notin \mathcal{X}_0, \end{cases}$$

and analogously define $P_{X|Y=y}(dx)$. Integrating these kernels against their respective marginals recovers P, so they are versions of the corresponding conditional laws. In row i of the batch loss, the anchor X_i is compared with $Y_1, \dots, Y_K$. We refer to these as the row candidates and to Y_i as the paired candidate.

Lemma A.5. *Fix $x \in \mathcal{X}_0$. Let J be uniform on $\{1, \dots, K\}$. Conditional on $J = j$, let $\check{Y}_j$ have law $P_{Y|X=x}$, and let $(\check{Y}_\ell)_{\ell \neq j}$ be independent draws from P_Y, independent of $\check{Y}_j$. Write $\check{Y}_{1:K} = (\check{Y}_1, \dots, \check{Y}_K)$. Then, for every nonnegative measurable F,*

$$\mathbb{E}_x\big[F(J, \check{Y}_{1:K})\big] = \frac{1}{K} \sum_{j=1}^{K} \int F(j, y_{1:K}) r^\star(x, y_j) P_Y^{\otimes K}(dy_{1:K}), \tag{15}$$

where $\mathbb{P}_x$, $\mathbb{E}_x$, and Var_x denote probability, expectation, and variance under this construction. Consequently, the marginal law of $\check{Y}_{1:K}$ has density

$$m_x(y_{1:K}) = \frac{1}{K} \sum_{j=1}^{K} r^\star(x, y_j) \tag{16}$$

with respect to $P_Y^{\otimes K}$. Define

$$\check{\pi}_x^\star(j) = \frac{r^\star(x, \check{Y}_j)}{\sum_{\ell=1}^{K} r^\star(x, \check{Y}_\ell)}.$$

Then

$$\mathbb{P}_x(J = j \mid \check{Y}_{1:K}) = \check{\pi}_x^\star(j) \qquad \text{almost surely.} \tag{17}$$

Finally, let $\Phi : \{1, \dots, K\} \times \mathcal{Y}^K \to \mathbb{R}$ be measurable and invariant under simultaneously permuting the candidate coordinates and the paired candidate index. Assume that Φ is nonnegative or that $\mathbb{E}_x|\Phi(J, \check{Y}_{1:K})| < \infty$. Then, for every row index i,

$$\mathbb{E}\big[\Phi(i, Y_{1:K}) \,\big|\, X_i = x\big] = \mathbb{E}_x\big[\Phi(J, \check{Y}_{1:K})\big].$$

Proof. Conditional on $J = j$, the density of $\check{Y}_{1:K}$ with respect to $P_Y^{\otimes K}$ is $r^\star(x, y_j)$, which proves (15). Summing over j gives (16), and Bayes' formula gives (17).

Conditional on $J = j$, the vector $\check{Y}_{1:K}$ has the same law as $Y_{1:K}$ conditional on $X_i = x$, up to a permutation that moves the paired candidate from position i to position j. The final identity then follows from the assumed permutation invariance of Φ. □

Using the construction above, the conditional excess cross-entropy can be written as a KL divergence. The equality case then forces the score difference $\Delta = s - s^\star$ to be constant along each row and column, leading to the characterization of the population InfoNCE minimizers.

Proof of Proposition 3.1. Using (1) and (11), the batch loss can be written as

$$\ell_K(s; B) = \frac{1}{2K} \sum_{i=1}^{K} \{-\log \pi_i^s(i)\} + \frac{1}{2K} \sum_{j=1}^{K} \{-\log \bar{\pi}_j^s(j)\} - \log K.$$

Now we first show that

$$\mathcal{E}_K(s) = \frac{1}{2K} \mathbb{E}\left[\sum_{i=1}^{K} D_{\mathrm{KL}}(\pi_i^\star \| \pi_i^s) + \sum_{j=1}^{K} D_{\mathrm{KL}}(\bar{\pi}_j^\star \| \bar{\pi}_j^s) \right]. \tag{18}$$

Fix i and condition on $X_i = x$. By Lemma A.5, writing $\check{\pi}_x^s(j) = e^{s(x, \check{Y}_j)} / \sum_{\ell=1}^{K} e^{s(x, \check{Y}_\ell)}$, we have $\mathbb{E}[-\log \pi_i^s(i) \mid X_i = x] = \mathbb{E}_x[-\log \check{\pi}_x^s(J)]$. Conditioning further on $\check{Y}_{1:K}$ and using (17),

$$\mathbb{E}_x\left[-\log \check{\pi}_x^s(J) \,\middle|\, \check{Y}_{1:K}\right] = \sum_{j=1}^{K} \check{\pi}_x^\star(j) \{-\log \check{\pi}_x^s(j)\}. \tag{19}$$

For $s = s^\star$, the right-hand side is the entropy of $\check{\pi}^\star_x$ and hence lies in $[0, \log K]$; in particular, $L_K(s^\star) < \infty$. Subtracting (19) for $s^\star$ from the same identity for s gives

$$\mathbb{E}\big[-\log \pi_i^s(i) + \log \pi_i^{s^\star}(i) \mid X_i = x\big] = \mathbb{E}_x D_{\mathrm{KL}}\big(\check{\pi}^\star_x \,\big\|\, \check{\pi}^s_x\big). \tag{20}$$

The KL divergence between the two softmax vectors is invariant under permutations of the candidate coordinates. Hence Lemma A.5 implies that the expectation on the right-hand side of (20) equals the expected KL divergence in the original row. Integrating over $x \sim P_X$,

$$\mathbb{E}\big[-\log \pi_i^s(i) + \log \pi_i^{s^\star}(i)\big] = \mathbb{E} D_{\mathrm{KL}}(\pi_i^\star \| \pi_i^s).$$

The same argument applies to the columns. Summing the row and column identities proves (18).

Since the right-hand side of (18) is nonnegative, $s^\star$ is a minimizer of L_K. Now let s be any minimizer and set $\Delta = s - s^\star$. Then every KL divergence in (18) vanishes almost surely. Since all softmax probabilities are positive, equality of the two softmax distributions in the first row implies $\Delta(X_1, Y_1) = \Delta(X_1, Y_2)$ almost surely. Similarly, equality in the first column gives $\Delta(X_1, Y_1) = \Delta(X_2, Y_1)$ almost surely.

We use the elementary fact that if two independent random variables are equal almost surely, then both are almost surely constant. Conditioning on $X_1 = x$, for P_X-almost every x, $\Delta(x, Y_1) = \Delta(x, Y_2)$ almost surely, where $Y_1 \sim P_{Y|X=x}$ and $Y_2 \sim P_Y$ are independent. Hence $\Delta(x, \cdot)$ is P_Y-almost surely equal to a constant $f(x)$, so $\Delta(X, Y) = f(X)$ Q-almost surely. Symmetrically, $\Delta(X, Y) = g(Y)$ Q-almost surely for some function g. Since X and Y are independent under Q, the equality $f(X) = g(Y)$ almost surely implies that both sides are almost surely equal to a constant. Thus $s = s^\star + c$ Q-almost surely for some $c \in \mathbb{R}$.

Conversely, $L_K(s^\star + c) = L_K(s^\star)$ for every $c \in \mathbb{R}$ by the additive invariance of the InfoNCE loss, completing the proof. □

A.3 Finite-batch comparison

We next compare the finite-batch variance functional V_K with its infinite-batch version,

$$V_\infty(\Delta) = \frac{1}{2}\left[\mathbb{E}_X \mathrm{Var}_{Y|X}\{\Delta(X, Y)\} + \mathbb{E}_Y \mathrm{Var}_{X|Y}\{\Delta(X, Y)\}\right].$$

The following identity expresses $V_K(\Delta)$ as $V_\infty(\Delta)$ minus a finite-batch correction. As shown in the proof, the correction is the variance of a posterior mean and is therefore nonnegative, implying in particular that $V_K(\Delta) \le V_\infty(\Delta)$.

Lemma A.6. *Let Δ be a bounded perturbation. For $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, define the centered row and column sections by*

$$\Delta_x^{\mathrm{row}}(y') = \Delta(x, y') - \int \Delta(x, v) P_{Y|X=x}(dv),$$
$$\Delta_y^{\mathrm{col}}(x') = \Delta(x', y) - \int \Delta(u, y) P_{X|Y=y}(du).$$

Let $\widetilde{Y}_{1:K} = (\widetilde{Y}_1, \dots, \widetilde{Y}_K) \sim P_Y^{\otimes K}$ independently of $X \sim P_X$, and let $\widetilde{X}_{1:K} = (\widetilde{X}_1, \dots, \widetilde{X}_K) \sim P_X^{\otimes K}$ independently of $Y \sim P_Y$. Then,

$$V_K(\Delta) = V_\infty(\Delta) - \frac{1}{2K}\mathbb{E}_X \mathbb{E}_{\widetilde{Y}_{1:K}} \frac{\left\{\sum_{j=1}^K r^\star(X, \widetilde{Y}_j) \Delta_X^{\mathrm{row}}(\widetilde{Y}_j)\right\}^2}{\sum_{j=1}^K r^\star(X, \widetilde{Y}_j)}$$
$$- \frac{1}{2K}\mathbb{E}_Y \mathbb{E}_{\widetilde{X}_{1:K}} \frac{\left\{\sum_{i=1}^K r^\star(\widetilde{X}_i, Y) \Delta_Y^{\mathrm{col}}(\widetilde{X}_i)\right\}^2}{\sum_{i=1}^K r^\star(\widetilde{X}_i, Y)}. \tag{21}$$

Proof. Step 1. We first prove the identity for the row half of V_K and V_∞. In the row calculations, we abbreviate $\Delta_x = \Delta_x^{\mathrm{row}}$. For $x \in \mathcal{X}_0$,

$$\int r^\star(x, y) \Delta_x(y) P_Y(dy) = 0.$$

The K rows of a batch have the same distribution, so

$$\frac{1}{2K}\mathbb{E}\sum_{i=1}^{K}\mathrm{Var}_{\pi_i^\star}\big(\Delta(X_i,Y_1),\dots,\Delta(X_i,Y_K)\big)$$
$$=\frac{1}{2}\mathbb{E}_X\,\mathbb{E}\big[\mathrm{Var}_{\pi_1^\star}\big(\Delta(X_1,Y_1),\dots,\Delta(X_1,Y_K)\big)\,\big|\,X_1=X\big],$$

and it suffices to show that, for P_X-almost every x,

$$\mathbb{E}\big[\mathrm{Var}_{\pi_1^\star}\big(\Delta(x,Y_1),\dots,\Delta(x,Y_K)\big)\,\big|\,X_1=x\big]$$
$$=\mathrm{Var}_{Y|X=x}\{\Delta(x,Y)\}-\frac{1}{K}\mathbb{E}_{\widetilde{Y}_{1:K}}\frac{\big\{\sum_{j=1}^K r^\star(x,\widetilde{Y}_j)\Delta_x(\widetilde{Y}_j)\big\}^2}{\sum_{j=1}^K r^\star(x,\widetilde{Y}_j)}. \tag{22}$$

Here, conditional on $X_1=x$, the first candidate Y_1 has density $r^\star(x,\cdot)$ with respect to P_Y. We fix such an x and use the construction introduced in Lemma A.5.

Step 2. For $y_{1:K}=(y_1,\dots,y_K)$, define

$$\pi_x^\star(j;y_{1:K})=\frac{r^\star(x,y_j)}{\sum_{\ell=1}^K r^\star(x,y_\ell)},\qquad j=1,\dots,K.$$

Write

$$g(y_{1:K})=\mathrm{Var}_{\pi_x^\star(\,\cdot\,;y_{1:K})}\big(\Delta(x,y_1),\dots,\Delta(x,y_K)\big).$$

Permuting the candidates permutes the weights $\pi_x^\star(j;y_{1:K})$ and the values $\Delta(x,y_j)$ together, so g is a symmetric function of $y_{1:K}$. Given $J=j$, the vector $\check{Y}_{1:K}$ has the same distribution as the batch row conditional on $X_1=x$, after moving its paired candidate from position 1 to position j. By symmetry of g,

$$\mathbb{E}_x\{g(\check{Y}_{1:K})\mid J=j\}=\mathbb{E}\{g(Y_{1:K})\mid X_1=x\}\qquad\text{for every } j,$$

and averaging over J gives $\mathbb{E}_x g(\check{Y}_{1:K})=\mathbb{E}\{g(Y_{1:K})\mid X_1=x\}$. On the other hand, by (17), conditional on $\check{Y}_{1:K}$, the value $\Delta(x,\check{Y}_J)$ equals $\Delta(x,\check{Y}_j)$ with probability $\check{\pi}_x^\star(j)$, so

$$g(\check{Y}_{1:K})=\mathrm{Var}_x\big\{\Delta(x,\check{Y}_J)\,\big|\,\check{Y}_{1:K}\big\}.$$

Combining the last two equations, the left-hand side of (22) equals $\mathbb{E}_x\mathrm{Var}_x\big\{\Delta(x,\check{Y}_J)\,\big|\,\check{Y}_{1:K}\big\}$.

Step 3. Under this construction, $\check{Y}_J$ has law $P_{Y|X=x}$. Hence $\mathrm{Var}_x\{\Delta(x,\check{Y}_J)\}=\mathrm{Var}_{Y|X=x}\{\Delta(x,Y)\}$. The law of total variance then gives

$$\mathrm{Var}_{Y|X=x}\{\Delta(x,Y)\}=\mathbb{E}_x\mathrm{Var}_x\big\{\Delta(x,\check{Y}_J)\,\big|\,\check{Y}_{1:K}\big\}$$
$$+\,\mathrm{Var}_x\left[\mathbb{E}_x\big\{\Delta(x,\check{Y}_J)\,\big|\,\check{Y}_{1:K}\big\}\right]. \tag{23}$$

The first term on the right is the row variance from Step 2. The identity (22) therefore follows once we show that the second term, the variance of the posterior mean, equals the subtracted term in (22).

Step 4. By (17), the posterior mean is

$$\mathbb{E}_x\big\{\Delta(x,\check{Y}_J)\,\big|\,\check{Y}_{1:K}\big\}=\sum_{j=1}^K\check{\pi}_x^\star(j)\Delta(x,\check{Y}_j).$$

Its expectation under this construction is $\mathbb{E}_x\Delta(x,\check{Y}_J)=\mathbb{E}\{\Delta(x,Y)\mid X=x\}$, the centering constant in the definition of Δ_x. Since the weights $\check{\pi}_x^\star(j)$ sum to one,

$$\sum_{j=1}^K\check{\pi}_x^\star(j)\Delta(x,\check{Y}_j)-\mathbb{E}\{\Delta(x,Y)\mid X=x\}$$
$$=\sum_{j=1}^K\check{\pi}_x^\star(j)\Delta_x(\check{Y}_j)=\frac{T}{S},$$

where $S = \sum_{j=1}^{K} r^\star(x, \check{Y}_j)$ and $T = \sum_{j=1}^{K} r^\star(x, \check{Y}_j)\Delta_x(\check{Y}_j)$. Hence the variance of the posterior mean is $\mathbb{E}_x\{(T/S)^2\}$. By Lemma A.5, the law of $\check{Y}_{1:K}$ has density $m_x = S/K$ with respect to $P_Y^{\otimes K}$. Changing measure to $P_Y^{\otimes K}$, or equivalently replacing $\check{Y}_{1:K}$ by independent draws $\widetilde{Y}_{1:K}$ from P_Y, gives

$$\mathrm{Var}_x\big[\mathbb{E}_x\{\Delta(x, \check{Y}_J) \,\big|\, \check{Y}_{1:K}\}\big] = \mathbb{E}_{\widetilde{Y}_{1:K}}\left[\frac{\widetilde{S}}{K}\left(\frac{\widetilde{T}}{\widetilde{S}}\right)^2\right] = \frac{1}{K}\,\mathbb{E}_{\widetilde{Y}_{1:K}}\frac{\widetilde{T}^2}{\widetilde{S}},$$

where $\widetilde{S}$ and $\widetilde{T}$ are S and T with $\widetilde{Y}_j$ in place of $\check{Y}_j$. This is exactly the subtracted term in (22).

Step 5. Substituting Step 4 into (23) and using Step 2 proves (22). Averaging over $x \sim P_X$ with the factor $1/2$, and repeating the argument for columns with $Y = y$ fixed, gives (21). Both subtracted terms are expectations of nonnegative quantities, so $V_K(\Delta) \le V_\infty(\Delta)$. □

Since $\|s^\star\|_\infty \le M_0$ by Assumption 3.2 and $r^\star = e^{s^\star}$, we have

$$r_{\min} \le r^\star \le r_{\max} \quad Q\text{-almost surely}, \qquad r_{\min} = e^{-M_0}, \quad r_{\max} = e^{M_0}. \tag{24}$$

The density-ratio bounds also give lower bounds on the finite-batch curvature. The next proposition provides a comparison and a lower bound in terms of the variance under the product distribution.

Proposition A.7. *Under (24), every perturbation Δ with $\|\Delta\|_\infty < \infty$ satisfies, for every $K \ge 2$,*

$$\max\left\{\frac{K-1}{K r_{\max}}, \left(1 - \frac{r_{\max}}{r_{\min} K}\right)_+\right\} V_\infty(\Delta) \le V_K(\Delta) \le V_\infty(\Delta), \qquad V_\infty(\Delta) \ge \frac{r_{\min}}{2}\mathrm{Var}_Q(\Delta).$$

Proof. Fix $x \in \mathcal{X}_0$ and abbreviate, for independent draws $\widetilde{Y}_1, \dots, \widetilde{Y}_K$ from P_Y,

$$w_j = r^\star(x, \widetilde{Y}_j), \qquad h_j = \Delta_x(\widetilde{Y}_j), \qquad S = \sum_{j=1}^{K} w_j, \qquad \pi_j = \frac{w_j}{S}.$$

By Steps 1 and 2 of the proof of Lemma A.6, the row half of V_K conditioned on the anchor x is the expectation of the row softmax variance under the construction of Lemma A.5, and the change of measure (16) converts it into an expectation under the product law:

$$\mathbb{E}\big[\mathrm{Var}_{\pi_1^\star}\big(\Delta(x, Y_1), \dots, \Delta(x, Y_K)\big) \,\big|\, X_1 = x\big] = \frac{1}{K}\,\mathbb{E}\big\{S\,\mathrm{Var}_\pi(h_1, \dots, h_K)\big\},$$

where we also used that the softmax variance is unchanged when the constant $\mathbb{E}\{\Delta(x, Y) \mid X = x\}$ is subtracted from every value. Now expand the weighted variance:

$$S\,\mathrm{Var}_\pi(h) = \sum_j w_j h_j^2 - \frac{\big(\sum_j w_j h_j\big)^2}{S} = \frac{1}{S}\sum_{j<\ell} w_j w_\ell (h_j - h_\ell)^2.$$

Every term is nonnegative, and $S \le K r_{\max}$ because $w_j \le r_{\max}$, so

$$\frac{1}{K}\,\mathbb{E}\big\{S\,\mathrm{Var}_\pi(h)\big\} \ge \frac{1}{K^2 r_{\max}}\sum_{j<\ell}\mathbb{E}\big\{w_j w_\ell (h_j - h_\ell)^2\big\}.$$

For $j \ne \ell$, the draws are independent, and since $x \in \mathcal{X}_0$ and Δ_x is centered under $P_{Y|X=x}$ (Step 1 of the proof of Lemma A.6), we have $\mathbb{E}w_j = 1$ and $\mathbb{E}(w_j h_j) = 0$, so

$$\mathbb{E}\big\{w_j w_\ell (h_j - h_\ell)^2\big\} = 2\,\mathbb{E}(w_1 h_1^2) = 2\mathrm{Var}_{Y|X=x}\{\Delta(x, Y)\},$$

where the last equality uses that $r^\star(x, \cdot)P_Y = P_{Y|X=x}$ and that Δ_x is centered under this law. Thus

$$\mathbb{E}\big[\mathrm{Var}_{\pi_1^\star}\big(\Delta(x, Y_1), \dots, \Delta(x, Y_K)\big) \,\big|\, X_1 = x\big] \ge \frac{K-1}{K r_{\max}}\mathrm{Var}_{Y|X=x}\{\Delta(x, Y)\}.$$

Multiplying by $1/2$, integrating over $x \sim P_X$, and repeating the argument for columns gives $V_K(\Delta) \ge \{(K-1)/(K r_{\max})\}V_\infty(\Delta)$.

With $T = \sum_j w_j h_j$ and S as above, $S \geq K r_{\min}$, so $T^2/S \leq T^2/(K r_{\min})$. The summands of T are independent with mean zero, so

$$\begin{aligned}\mathbb{E}T^2 = K \int r^\star(x,y)^2 \Delta_x(y)^2 P_Y(dy) &\leq K r_{\max} \int r^\star(x,y) \Delta_x(y)^2 P_Y(dy) \\ &= K r_{\max} \mathrm{Var}_{Y|X=x}\{\Delta(x,Y)\}.\end{aligned}$$

Therefore

$$\frac{1}{K}\mathbb{E}\frac{T^2}{S} \leq \frac{r_{\max}}{r_{\min} K} \mathrm{Var}_{Y|X=x}\{\Delta(x,Y)\}.$$

Multiplying by $1/2$, integrating over x, and adding the column term shows that the sum of the two subtracted terms in (21) is at most $\{r_{\max}/(r_{\min}K)\}V_\infty(\Delta)$, so the identity gives $V_K \geq \{1 - r_{\max}/(r_{\min}K)\}_+ V_\infty$.

To prove $V_\infty(\Delta) \geq (r_{\min}/2)\mathrm{Var}_Q(\Delta)$, fix x and a bounded function f of y. Since $P_{Y|X=x}$ has density $r^\star(x,\cdot) \geq r_{\min}$ with respect to P_Y,

$$\begin{aligned}\mathrm{Var}_{Y|X=x}\{f(Y)\} &= \inf_{c\in\mathbb{R}} \int \{f(y) - c\}^2 r^\star(x,y) P_Y(dy) \\ &\geq r_{\min} \inf_{c\in\mathbb{R}} \int \{f(y) - c\}^2 P_Y(dy) = r_{\min} \mathrm{Var}_{P_Y}(f).\end{aligned}$$

Applying this with $f = \Delta(x,\cdot)$, and the analogous bound for columns,

$$V_\infty(\Delta) \geq \frac{r_{\min}}{2}\left[\mathbb{E}_X \mathrm{Var}_{P_Y}\{\Delta(X,\cdot)\} + \mathbb{E}_Y \mathrm{Var}_{P_X}\{\Delta(\cdot,Y)\}\right].$$

Finally, under the product law Q the variance of Δ decomposes as

$$\mathrm{Var}_Q(\Delta) = \mathbb{E}_X \mathrm{Var}_{P_Y}\{\Delta(X,\cdot)\} + \mathrm{Var}_{P_X}\left[\mathbb{E}_{P_Y}\{\Delta(X,\cdot)\}\right],$$

and the variance of the average $\mathbb{E}_{P_Y}\{\Delta(\cdot,Y)\}$ is at most the average of the variances, $\mathbb{E}_Y \mathrm{Var}_{P_X}\{\Delta(\cdot,Y)\}$, by the Cauchy-Schwarz inequality. Hence the bracket in the above equation is at least $\mathrm{Var}_Q(\Delta)$, which gives $V_\infty(\Delta) \geq (r_{\min}/2)\mathrm{Var}_Q(\Delta)$. □

Combining the lower Taylor estimate with the finite-batch comparison proves Proposition 4.1. The resulting constant is independent of the batch size.

Proof of Proposition 4.1. Write $\Delta = s - s^\star$. Proposition A.3 gives $\mathcal{E}_K(s) \geq V_K(\Delta)/(\kappa + 2)$. Proposition A.7 gives

$$V_K(\Delta) \geq \frac{K-1}{K r_{\max}} V_\infty(\Delta) \geq \frac{1}{2 r_{\max}} V_\infty(\Delta), \qquad V_\infty(\Delta) \geq \frac{r_{\min}}{2}\mathrm{Var}_Q(\Delta),$$

where the second inequality uses $K \geq 2$. Using $\mathrm{Var}_Q(\Delta) = \inf_c \|\Delta - c\|^2_{L^2(Q)} = \mathrm{Err}_Q(s)$ proves the claim. Thus we may take

$$C_1 = \frac{4(\kappa+2) r_{\max}}{r_{\min}} = 4(4M_0 + 2)e^{2M_0} \geq 1.$$

This constant depends only on M_0 and is independent of K. □

B Concentration and the oracle inequality

In this section, we prove Theorem 4.2, which establishes an oracle inequality for the empirical InfoNCE minimizer. We recall that $\kappa = 4M_0$ as defined in Appendix A.

B.1 SENSITIVITY TO PAIR REPLACEMENTS

We now turn to controlling deviations of the empirical loss from its expectation. The first step is to control the effect of replacing one or two paired observations within a batch. For these calculations, it is convenient to write the loss as

$$\ell_K(s;B) = \frac{1}{2K}\sum_{i=1}^{K}\left[\log\left(\frac{1}{K}\sum_{j=1}^{K} e^{s(X_i,Y_j)}\right) + \log\left(\frac{1}{K}\sum_{j=1}^{K} e^{s(X_j,Y_i)}\right) - 2s(X_i,Y_i)\right]. \quad (25)$$

Recall that the population excess risk of a score s is $\mathcal{E}_K(s) = L_K(s) - L_K(s^\star)$. The approximation error is $\inf_{s\in\mathcal{S}} \mathcal{E}_K(s)$, which is zero when the class $\mathcal{S}$ attains the population optimum, up to an additive constant. Write

$$W_s = \frac{1}{N}\sum_{b=1}^{N}\big\{\ell_K(s;B_b) - \ell_K(s^\star;B_b)\big\} - \mathcal{E}_K(s),$$

and, for any two scores s, t, we define

$$h_{s,t} = \ell_K(s;\cdot) - \ell_K(t;\cdot), \qquad \text{so that} \qquad W_s - W_t = \frac{1}{N}\sum_{b=1}^{N} h_{s,t}(B_b) - \mathbb{E}\, h_{s,t}(B).$$

Whenever two scores s, t are compared, write

$$\Delta_{s,t} = s - t - Q(s-t), \qquad d_Q(s,t) = \|\Delta_{s,t}\|_{L^2(Q)}, \qquad d_\infty(s,t) = 2\|s-t\|_\infty,$$

(here, $Qh = \int h dQ$). When s, t are fixed, we simply write d_Q and d_∞. Adding a constant to either score does not change $h_{s,t}$ or $W_s - W_t$. The values of $\Delta_{s,t}$ lie in an interval of length at most d_∞, and

$$\|\Delta_{s,t}\|_\infty \le d_\infty \le 4M_0, \qquad d_Q^2 = \mathrm{Var}_Q(s-t) \le Q\big((s-t)^2\big) \le \|s-t\|_\infty^2 = \frac{d_\infty^2}{4}.$$

Also, the evaluations of $\Delta_{s,t}$ that appear below are always of the form $\Delta_{s,t}(X_i, Y_j)$: for $i \neq j$ the pair (X_i, Y_j) has law Q, and for $i = j$ it has law $P = r^\star Q$; hence

$$\mathbb{E}\,\Delta_{s,t}(X_i,Y_j)^2 \le r_{\max} d_Q^2 \qquad \text{for every } i, j, \quad (26)$$

and the same bound holds when either coordinate is replaced by an independent copy, since the pair then again has law Q or P.

To control the effect of replacing one or two pairs in a batch, we first derive a representation of the loss difference in terms of softmax-weighted means of the centered difference $\Delta_{s,t}$. Fix a batch B and two scores s, t, and interpolate along the segment $s_\theta := (1-\theta)t + \theta s$ for $\theta \in [0,1]$; note $\|s_\theta\|_\infty \le M_0$ for every θ. Let

$$\pi_i^\theta(j) = \frac{e^{s_\theta(X_i,Y_j)}}{\sum_{\ell=1}^{K} e^{s_\theta(X_i,Y_\ell)}}, \qquad \bar\pi_j^\theta(i) = \frac{e^{s_\theta(X_i,Y_j)}}{\sum_{\ell=1}^{K} e^{s_\theta(X_\ell,Y_j)}}$$

be the row and column softmax weights of the score s_θ, as given in (11). Differentiating the expanded form (25) of the loss with respect to θ, and using that adding the constant $-Q(s-t)$ to the direction changes no weighted mean of differences, we obtain

$$h_{s,t}(B) = \int_0^1 \left[\frac{1}{2K}\sum_{i=1}^{K}\Big\{M_i(\theta) - \Delta_{s,t}(X_i,Y_i)\Big\} + \frac{1}{2K}\sum_{j=1}^{K}\Big\{\bar M_j(\theta) - \Delta_{s,t}(X_j,Y_j)\Big\}\right] d\theta, \quad (27)$$

where

$$M_i(\theta) = \sum_{j=1}^{K} \pi_i^\theta(j)\,\Delta_{s,t}(X_i,Y_j), \qquad \bar M_j(\theta) = \sum_{i=1}^{K} \bar\pi_j^\theta(i)\,\Delta_{s,t}(X_i,Y_j),$$

are the row and column softmax means of the $\Delta_{s,t}$-values.

The first consequence is a deterministic Lipschitz bound in the sup norm. Besides providing a crude uniform bound, it will later control the approximation error when the score class is replaced by a finite net in the chaining argument.

Lemma B.1. *For all scores s, t bounded by M_0, the following bounds hold pointwise for every batch:*

$$|h_{s,t}(B)| \le d_\infty(s,t), \qquad |W_s - W_t| \le 2d_\infty(s,t).$$

Proof. The values of $\Delta_{s,t}$ lie in an interval of length at most $d_\infty = 2\|s-t\|_\infty$. Every weighted mean and its corresponding diagonal value belong to this interval. Each brace in the integral representation of $h_{s,t}$ therefore has absolute value at most d_∞, so $|h_{s,t}(B)| \le d_\infty$. Both its empirical and population averages also have absolute value at most d_∞, which gives $|W_s - W_t| \le 2d_\infty$. □

Taking expectations in Lemma B.1 and using Proposition 3.1 gives

$$0 \le \mathcal{E}_K(s) \le 2\|s - s^\star\|_\infty \le 4M_0 = \kappa, \qquad s \in \mathcal{S}. \tag{28}$$

Consequently, by the definition of $\mathcal{S}(\rho)$, we have $\mathcal{S} = \mathcal{S}(\sqrt{\kappa})$.

The next lemma considers a generic weighted mean of the form appearing in (27) and quantifies how much it can change when one or two of its atoms are replaced. The argument is entirely deterministic and does not use independence among the atoms.

Lemma B.2. *Let $\Lambda \ge 1$. Let $w_1, \dots, w_K > 0$ be weights, let $u_1, \dots, u_K$ be real values, and set*

$$S = \frac{1}{K}\sum_{j=1}^K w_j, \qquad A = \frac{1}{K}\sum_{j=1}^K w_j u_j, \qquad M = \frac{A}{S} = \sum_{j=1}^K \frac{w_j}{KS}\, u_j.$$

For a replacement of atom m, let $M^{(m)}$ denote the weighted mean obtained by replacing (w_m, u_m) with (w'_m, u'_m), and define $D_m M = M^{(m)} - M$. For distinct m, ℓ, let $M^{(m,\ell)}$ denote the weighted mean obtained by replacing both (w_m, u_m) and (w_ℓ, u_ℓ) with (w'_m, u'_m) and (w'_ℓ, u'_ℓ), and define $D_\ell D_m M = M^{(m,\ell)} - M^{(m)} - M^{(\ell)} + M$. Assume

$$\frac{\max\{w_1, \dots, w_K, w'_m, w'_\ell\}}{\min\{w_1, \dots, w_K, w'_m, w'_\ell\}} \le \Lambda.$$

(i) M is a weighted average of $u_1, \dots, u_K$ with weights $w_j/(KS) \le \Lambda/K$; in particular $\min_j u_j \le M \le \max_j u_j$ and $M^2 \le (\Lambda/K)\sum_j u_j^2$.

(ii) With $\delta A = (w'_m u'_m - w_m u_m)/K$ and $\delta S = (w'_m - w_m)/K$, the exact identities

$$D_m M = \frac{\delta A - M\,\delta S}{S + \delta S}, \qquad \delta A - M\,\delta S = \frac{1}{K}\big\{w'_m(u'_m - M) - w_m(u_m - M)\big\}$$

hold, and consequently

$$|D_m M| \le \frac{\Lambda}{K}\big(|u'_m - M| + |u_m - M|\big), \qquad (D_m M)^2 \le \frac{2\Lambda^2}{K^2}\big\{(u'_m - M)^2 + (u_m - M)^2\big\}.$$

(iii) For $\ell \ne m$, with the increments $(\delta A_m, \delta S_m)$ and $(\delta A_\ell, \delta S_\ell)$ of the two replacements,

$$D_\ell D_m M = \int_0^1 \int_0^1 \left\{ -\frac{\delta A_m\,\delta S_\ell + \delta A_\ell\,\delta S_m}{\widetilde{S}^2} + \frac{2\widetilde{A}\,\delta S_m\,\delta S_\ell}{\widetilde{S}^3} \right\} d\sigma d\tau,$$

where $\widetilde{A} = A + \sigma\delta A_m + \tau\delta A_\ell$ and $\widetilde{S} = S + \sigma\delta S_m + \tau\delta S_\ell$. Consequently, if all the values u_j, u'_m, u'_ℓ are bounded in absolute value by a, then

$$|D_\ell D_m M| \le \frac{6\Lambda^3}{K^2}\, a.$$

Proof. All three quantities M, $D_m M$, $D_\ell D_m M$ are unchanged when the whole pool $w_1, \dots, w_K, w'_m, w'_\ell$ is multiplied by a common constant. We may therefore divide all weights by the smallest weight, after which every weight lies in $[1, \Lambda]$. In particular $S \ge 1$, and the same holds for the sum $S + \delta S$ of any replaced configuration and, in (iii), for $\widetilde{S}$: being affine in each of

σ, τ separately, $\widetilde{S}$ attains its minimum over the square at a corner, and every corner is an average of pool weights, hence at least 1.

(i) is immediate: the weights $w_j/(KS)$ are nonnegative, sum to one, and are bounded by $\Lambda/(K \cdot 1)$; the second bound follows from Jensen's inequality applied to this average.

(ii) The first identity is algebra: $D_m M = (A + \delta A)/(S + \delta S) - A/S = \{\delta A\, S - A\, \delta S\}/\{S(S + \delta S)\} = (\delta A - M\delta S)/(S + \delta S)$. The second follows by inserting $w'_m u'_m - w_m u_m = w'_m(u'_m - M) - w_m(u_m - M) + M(w'_m - w_m)$. For the bounds, $S + \delta S \geq 1$ and $w_m, w'_m \leq \Lambda$; squaring uses $(x + y)^2 \leq 2x^2 + 2y^2$.

(iii) Define $\varphi(\sigma, \tau) = \widetilde{A}/\widetilde{S}$ on $[0, 1]^2$. Because the two replaced atoms are distinct, the increments add: the four values $\varphi(1, 1), \varphi(1, 0), \varphi(0, 1), \varphi(0, 0)$ are exactly the four configurations of M, and $D_\ell D_m M = \varphi(1, 1) - \varphi(1, 0) - \varphi(0, 1) + \varphi(0, 0) = \int_0^1 \int_0^1 \partial_\sigma \partial_\tau \varphi d\sigma d\tau$; differentiating $\varphi = \widetilde{A}/\widetilde{S}$ twice gives the stated integrand. On the square, $\widetilde{S} \geq 1$ as shown above; moreover $|\delta S_m|, |\delta S_\ell| \leq (\Lambda - 1)/K \leq \Lambda/K$, $|\delta A_m|, |\delta A_\ell| \leq 2\Lambda a/K$, and $\left|\widetilde{A}\right| \leq \Lambda a$. Hence

$$|\partial_\sigma \partial_\tau \varphi| \leq 2 \cdot \frac{2\Lambda a}{K} \cdot \frac{\Lambda}{K} + 2 \cdot \Lambda a \cdot \frac{\Lambda^2}{K^2} = \frac{4\Lambda^2 + 2\Lambda^3}{K^2}\, a \leq \frac{6\Lambda^3}{K^2}\, a. \qquad \square$$

We now combine the atom-level bounds across the rows and columns of a batch. For a batch $B = (Z_1, \ldots, Z_K)$ with $Z_m = (X_m, Y_m)$ and a replacement Z'_m, let $B^{(m)}$ denote the batch obtained by replacing Z_m with Z'_m. For distinct m, ℓ, let $B^{(m,\ell)}$ denote the batch obtained after both replacements. For any real-valued batch functional F, define

$$D_m F(B) = F(B^{(m)}) - F(B),$$
$$D_\ell D_m F(B) = F(B^{(m,\ell)}) - F(B^{(m)}) - F(B^{(\ell)}) + F(B).$$

From now on, throughout this subsection, we write

$$\Lambda = e^{2M_0}.$$

The following lemma consists of three parts. The first controls the worst possible effect of one pair, the second keeps the much smaller $L^2(Q)$ size in the variance term, and the third shows that the reuse of negatives creates only a second-order interaction.

Lemma B.3. *For all scores $s, t \in \mathcal{S} \cup \{s^\star\}$, every $K \geq 2$, and all coordinates $m \neq \ell$:*

(a) $|D_m h_{s,t}| \leq \dfrac{C_2\, d_\infty}{K}$ *for every batch and every choice of the replacement pair;*

(b) $\mathbb{E}(D_m h_{s,t})^2 \leq \dfrac{C_2\, d_Q^2}{K^2}$ *when pair m is replaced by an independent copy;*

(c) $|D_\ell D_m h_{s,t}| \leq \dfrac{C_2\, d_\infty}{K^2}$ *for every batch and all choices of the replacement pairs.*

Here $d_\infty = 2\|s - t\|_\infty$ and $d_Q = d_Q(s, t)$ as defined earlier, and one may take $C_2 = 100\Lambda^3 r_{\max}$.

Proof. For a fixed $\theta \in [0, 1]$, let $\Phi_\theta(B)$ denote the integrand in (27). Then $D_m h_{s,t}(B) = \int_0^1 D_m \Phi_\theta(B) d\theta$, and similarly for $D_\ell D_m h_{s,t}$. Thus, any bound on $D_m \Phi_\theta$ or $D_\ell D_m \Phi_\theta$ that holds uniformly in θ can be integrated directly. For the second-moment estimate, Jensen's inequality gives $\mathbb{E}\{D_m h_{s,t}(B)\}^2 \leq \int_0^1 \mathbb{E}\{D_m \Phi_\theta(B)\}^2 d\theta$. It is therefore enough to prove all three estimates for fixed θ, with bounds uniform in θ.

We suppress the dependence on θ from the notation. The weights in row i are $w_j = e^{s_\theta(X_i, Y_j)}$. If $i \neq m$, replacing pair m changes only the atom at position m, while the anchor X_i remains fixed. Thus, all original and replacement weights involved in applying Lemma B.2 share the same anchor X_i, and

$$\frac{e^{s_\theta(X_i, y)}}{e^{s_\theta(X_i, y')}} \leq e^{2\|s_\theta\|_\infty} \leq e^{2M_0} = \Lambda.$$

The last inequality follows from $\|s_\theta\|_\infty \le (1-\theta)\|t\|_\infty + \theta\|s\|_\infty \le M_0$.

When $i = m$, the anchor itself changes, so we handle this row separately using the crude range bounds below. The column argument is identical with the roles of X and Y reversed.

(a) One pair, worst case. For a row $i \neq m$, part (ii) of Lemma B.2 with all values in $[-d_\infty, d_\infty]$ (recall $\|\Delta_{s,t}\|_\infty \le d_\infty$) gives, using $|\Delta_{s,t} - M| \le 2d_\infty$, the bound $|D_m(M_i - \Delta_{s,t}(X_i, Y_i))| = |D_m M_i| \le \frac{4\Lambda d_\infty}{K}$. For row m, both M_m and $\Delta_{s,t}(X_m, Y_m)$ lie in $[\inf \Delta_{s,t}, \sup \Delta_{s,t}]$, so the row term lies in $[-d_\infty, d_\infty]$ in both batches and its change is at most $2d_\infty$. Summing over the two row groups and doubling for the columns,

$$|D_m h_{s,t}| \le \frac{2}{2K}\Big\{(K-1)\frac{4\Lambda d_\infty}{K} + 2d_\infty\Big\} \le \frac{(4\Lambda+2)\, d_\infty}{K} \le \frac{6\Lambda d_\infty}{K}.$$

(b) One pair, second moment. We decompose the change into the contributions from the rows $i \neq m$, row m, the columns $j \neq m$, and column m. By the Cauchy-Schwarz inequality,

$$\mathbb{E}(D_m h_{s,t})^2 \le \frac{4}{(2K)^2}\Big\{\mathbb{E}\Big(\sum_{i\neq m} D_m M_i\Big)^2 + \mathbb{E}\big(D_m(M_m - \Delta_{s,t}(X_m, Y_m))\big)^2 + (\text{column analogues})\Big\}. \tag{29}$$

For a row $i \neq m$, part (ii) of Lemma B.2 gives

$$\begin{aligned}(D_m M_i)^2 &\le \frac{2\Lambda^2}{K^2}\big\{(\Delta_{s,t}(X_i, Y'_m) - M_i)^2 + (\Delta_{s,t}(X_i, Y_m) - M_i)^2\big\} \\ &\le \frac{4\Lambda^2}{K^2}\big\{\Delta_{s,t}(X_i, Y'_m)^2 + \Delta_{s,t}(X_i, Y_m)^2 + 2M_i^2\big\}.\end{aligned}$$

Both (X_i, Y_m) and (X_i, Y'_m) have law Q since $i \neq m$ and Y'_m is an independent copy, so their squared evaluations have mean d_Q^2. Moreover, by part (i) and (26), $\mathbb{E}M_i^2 \le (\Lambda/K)\sum_j \mathbb{E}\Delta_{s,t}(X_i, Y_j)^2 \le \Lambda r_{\max} d_Q^2$. Hence $\mathbb{E}(D_m M_i)^2 \le (8\Lambda^2 + 8\Lambda^3 r_{\max}) d_Q^2/K^2 \le 16\Lambda^3 r_{\max} d_Q^2/K^2$. Using $\big(\sum_{i\neq m} x_i\big)^2 \le (K-1)\sum_{i\neq m} x_i^2$,

$$\mathbb{E}\Big(\sum_{i\neq m} D_m M_i\Big)^2 \le (K-1)^2 \max_{i\neq m} \mathbb{E}(D_m M_i)^2 \le 16\Lambda^3 r_{\max} d_Q^2.$$

For row m, the anchor itself is replaced. The change is the difference between the row-m terms for B and $B^{(m)}$, so its square is at most

$$2(M_m - \Delta_{s,t}(X_m, Y_m))^2 + 2(M'_m - \Delta_{s,t}(X'_m, Y'_m))^2,$$

where M'_m is the row mean with anchor X'_m. Each squared row term has expectation at most $2\mathbb{E}M^2 + 2\mathbb{E}\Delta_{s,t}^2 \le 2\Lambda r_{\max} d_Q^2 + 2 r_{\max} d_Q^2 \le 4\Lambda r_{\max} d_Q^2$ by part (i) and (26). Thus, the row-m contribution is at most $16\Lambda r_{\max} d_Q^2$.

The same bounds hold for the corresponding column contributions. Substituting these estimates into (29) gives

$$\mathbb{E}(D_m h_{s,t})^2 \le \frac{1}{K^2}\cdot 2\big(16\Lambda^3 + 16\Lambda\big) r_{\max} d_Q^2 \le \frac{64\,\Lambda^3 r_{\max}\, d_Q^2}{K^2}.$$

(c) Two pairs, worst case. Fix $m \neq \ell$. We first consider the row contributions.

For a row $i \notin \{m, \ell\}$, replacing pairs m and ℓ changes only the two atoms at positions m and ℓ in M_i. The diagonal term $\Delta_{s,t}(X_i, Y_i)$ is unchanged, so its second difference is zero. Therefore, part (iii) of Lemma B.2 gives

$$|D_\ell D_m M_i| \le \frac{6\Lambda^3 d_\infty}{K^2}.$$

For row m, the diagonal term $\Delta_{s,t}(X_m, Y_m)$ does not depend on pair ℓ, so only the row mean M_m contributes after applying D_ℓ. With the anchor fixed, replacing pair ℓ changes only one atom of

M_m, and part (a) gives $|D_\ell M_m| \le 4\Lambda d_\infty/K$. This bound holds both before and after replacing pair m, and hence

$$|D_\ell D_m(\text{row-}m \text{ term})| \le \frac{8\Lambda d_\infty}{K}.$$

The same bound holds for row ℓ, and the column contributions are handled in the same way.

Summing the row and column contributions with the factor $1/(2K)$,

$$|D_\ell D_m h_{s,t}| \le \frac{2}{2K}\Big\{(K-2)\frac{6\Lambda^3 d_\infty}{K^2} + 2\cdot\frac{8\Lambda d_\infty}{K}\Big\} \le \frac{(6\Lambda^3+16\Lambda)d_\infty}{K^2} \le \frac{22\Lambda^3 d_\infty}{K^2}.$$

All three bounds therefore hold with $C_2 = 100\Lambda^3 r_{\max}$. □

B.2 Concentration of loss increments

This subsection combines the bounds from Lemma B.3 with concentration inequalities from Maurer (2019) to prove Proposition B.7. Although the paired observations are independent, the row and column contributions to the batch loss reuse the same observations and therefore are not independent summands.

For fixed scores s, t, recall that $d_\infty = 2\|s-t\|_\infty$ and $d_Q = d_Q(s,t)$. By Lemma B.3, replacing one paired observation changes the averaged loss difference by at most $C_2 d_\infty/n$. A direct application of McDiarmid's inequality (Boucheron et al., 2013, Theorem 6.2) therefore gives, with probability at least $1-2e^{-v}$, the deviation scale $d_\infty\sqrt{v/n}$, which does not exploit the localized distance d_Q. Alternatively, applying the classical bounded-summand Bernstein inequality (Boucheron et al., 2013, Theorem 2.10) to the N independent batch loss differences $h_{s,t}(B_1),\dots,h_{s,t}(B_N)$, whose centered absolute values are at most $2d_\infty$, gives the scale $d_Q\sqrt{\frac{v}{n}} + \frac{d_\infty v}{N}$. The linear term in this bound is not uniform in the batch size.

Maurer's inequality additionally uses the second-replacement bound in Lemma B.3. By Lemma B.3(b), the expected sum of conditional variances is of order d_Q^2/K, while the second-replacement bound controls the interaction term at order d_∞/K. Averaging over the independent batches then yields

$$d_Q\sqrt{\frac{v}{n}} + \frac{d_\infty v}{n},$$

as stated in Proposition B.7. Thus, the localized variance term is preserved without introducing an additional factor of K in the linear term.

For a function f of the K independent pairs, write $\mathbb{E}_m$ for the conditional expectation over the mth pair given the others, and

$$\Sigma^2(f) = \sum_{m=1}^{K} \mathbb{E}_m\big[(f-\mathbb{E}_m f)^2\big],$$

for the sum of conditional variances, so that $\mathbb{E}\,\Sigma^2(f) = \sum_m \mathbb{E}(f-\mathbb{E}_m f)^2$.

Lemma B.4. *Let $\mu = P^{\otimes K}$ be the batch product measure. For all scores $s,t \in \mathcal{S}\cup\{s^\star\}$ and every $K \ge 2$, the function $f := h_{s,t}(B)$ of the K independent pairs satisfies:*

(i) $\mathbb{E}\,\Sigma^2(f) \le \dfrac{C_2\, d_Q^2}{K}$;

(ii) $|f-\mathbb{E}_m f| \le \dfrac{C_2\, d_\infty}{K}$ *almost surely, for every m;*

(iii) Interaction functionals satisfy $J_\mu(f) \le J(f) \le \dfrac{C_2\, d_\infty}{K}$.

Here J and J_μ are the functionals defined in Maurer (2019, Section 1); all we use about them is

$$J_\mu(f)^2 \le J(f)^2 \le \sum_{k\neq\ell} \sup\ |D_\ell D_k f|^2, \tag{30}$$

where the supremum is over all batches and all choices of the replacement pairs.

Proof. We first prove (ii), and then establish (i) and (iii).

(ii) Since the replacement pair in $B^{(m)}$ is an independent copy, $\mathbb{E}_m f = \mathbb{E}[h_{s,t}(B^{(m)}) \mid B]$, so $f - \mathbb{E}_m f = -\mathbb{E}[D_m h_{s,t} \mid B]$. Lemma B.3(a) bounds its absolute value by $C_2 d_\infty / K$.

(i) By the same representation, Jensen's inequality, and Lemma B.3(b),

$$\mathbb{E}(f - \mathbb{E}_m f)^2 \le \mathbb{E}(D_m h_{s,t})^2 \le \frac{C_2 d_Q^2}{K^2}.$$

Summing over the K coordinates gives (i).

(iii) By (30) and Lemma B.3(c), which holds for every batch and all choices of the replacement pairs,

$$J(f)^2 \le K(K-1)\Big(\frac{C_2 d_\infty}{K^2}\Big)^2 \le \frac{C_2^2 d_\infty^2}{K^2}. \qquad \square$$

The Bernstein-type inequality of Maurer (2019, Theorem 1) states that a bounded measurable function f of independent variables with $f - \mathbb{E}_m f \le b$ for all m satisfies, for all $x > 0$,

$$\mathbb{P}(f - \mathbb{E}f > x) \le \exp\left(-\frac{x^2}{2\,\mathbb{E}\Sigma^2(f) + (2b/3 + J_\mu(f))\,x}\right).$$

Lemma B.4 gives a one-batch variance factor of order d_Q^2/K, which becomes small under localization, whereas d_∞ need not shrink. To preserve this localized variance when averaging the N independent batches, we use the moment generating function (mgf) bound underlying Maurer's inequality. This will allow both the variance factor and the interaction scale to gain an additional factor $1/N$.

For this purpose, we use the following consequence of Maurer's argument, which gives a sub-gamma mgf bound in terms of the expected conditional variance and the interaction functional.

Fact B.5. *Let f be a bounded measurable function of K independent random variables such that, for some $b, b_J > 0$,*

$$|f - \mathbb{E}_m f| \le b \quad \textit{a.s. for every } m, \qquad \textit{and} \qquad J_\mu(f) \le b_J.$$

Then, for both signs and all $0 \le \beta < (b/3 + b_J/2)^{-1}$,

$$\log \mathbb{E}\, e^{\pm\beta(f - \mathbb{E}f)} \le \frac{\beta^2\, \mathbb{E}\Sigma^2(f)/2}{1 - (b/3 + b_J/2)\,\beta}.$$

This statement is a direct consequence of the results in Maurer (2019); we give the derivation for completeness. By Maurer (2019, Section 2.1, Lemma 1),

$$\log \mathbb{E} e^{\beta(f - \mathbb{E}f)} = \beta \int_0^\beta \frac{S_f(\gamma)}{\gamma^2} d\gamma,$$

where S_f is the entropy-type quantity defined there. Moreover, the proof of Maurer (2019, Section 2.3, Proposition 3) shows that if $f - \mathbb{E}_m f \le 1$ for every m and $D(\Sigma^2(f)) \le \alpha^2 \Sigma^2(f)$, where D is the operator defined in Maurer (2019, Section 2.1, Definition 2), then

$$\int_0^\beta \frac{S_f(\gamma)}{\gamma^2} d\gamma \le \frac{\mathbb{E}\Sigma^2(f)}{2} \cdot \frac{\beta}{1 - (1/3 + \alpha/2)\beta}, \qquad 0 < \beta < \big(1/3 + \alpha/2\big)^{-1}.$$

Finally, Maurer (2019, Section 2.2, Proposition 2) gives $D(\Sigma^2(f)) \le J_\mu(f)^2 \Sigma^2(f)$. Combining these results proves Fact B.5 for the $+$ sign when $b = 1$ and $\alpha = b_J$. For general b, we apply this bound to f/b, noting that Σ^2 scales by b^{-2} and J_μ by b^{-1}. The $-$ sign follows by applying the same argument to $-f$.

The mgf bound in Fact B.5 applies to one batch. The next lemma shows how its variance factor and scale change when we average independent batches.

Lemma B.6. *Let $G_1, \dots, G_N$ be i.i.d. with $\mathbb{E}G_1 = 0$ and, for both signs and $0 \le \beta < 1/c_1$, $\log \mathbb{E}e^{\pm\beta G_1} \le (\beta^2 v_1/2)/(1 - c_1\beta)$. Then the average $\bar{G} := N^{-1}\sum_i G_i$ satisfies the same bound with parameters $(v_1/N,\, c_1/N)$, and consequently, for every $v \ge 0$,*

$$\mathbb{P}\left\{ |\bar{G}| > \sqrt{\frac{2v_1 v}{N}} + \frac{c_1 v}{N} \right\} \le 2e^{-v}.$$

Proof. By independence,

$$\log \mathbb{E}e^{\pm\beta\bar{G}} = N \log \mathbb{E}e^{\pm(\beta/N)G_1} \le N \cdot \frac{(\beta/N)^2 v_1/2}{1 - c_1\beta/N} = \frac{\beta^2 (v_1/N)/2}{1 - (c_1/N)\beta}, \qquad 0 \le \beta < \frac{N}{c_1}.$$

This random variable is sub-gamma with variance factor v_1/N and scale c_1/N, and the two-sided tail inequality in the lemma is the standard sub-gamma tail bound (Boucheron et al., 2013, Section 2.4). □

We now average the N independent batch loss differences. By Lemma B.4 and Fact B.5, the one-batch variance factor and scale are of orders d_Q^2/K and d_∞/K, respectively. By Lemma B.6, averaging the N independent batches reduces each by a factor $1/N$; since $n = NK$, they become of orders d_Q^2/n and d_∞/n.

Proposition B.7. *Under (24) there is a constant C_3 such that for every $K \ge 2$, $N \ge 1$, all $s, t \in \mathcal{S} \cup \{s^\star\}$, and all $v \ge 0$,*

$$\mathbb{P}\left\{ |W_s - W_t| > C_3 \left(d_Q(s,t)\sqrt{\frac{v}{n}} + d_\infty(s,t)\frac{v}{n} \right) \right\} \le 2e^{-v}.$$

Proof. If $d_\infty = 0$, then $s = t$ and the result is immediate. Assume $d_\infty > 0$. The batch loss differences $h_{s,t}(B_1), \dots, h_{s,t}(B_N)$ are independent copies of a function of K independent pairs which, by Lemma B.4, satisfies the hypotheses of Fact B.5 with

$$\mathbb{E}\Sigma^2(h_{s,t}) \le \frac{C_2 d_Q^2}{K} =: v_1, \qquad b = b_J = \frac{C_2 d_\infty}{K}, \qquad \text{so } \frac{b}{3} + \frac{b_J}{2} = \frac{5C_2 d_\infty}{6K} =: c_1.$$

By Lemma B.6 and $KN = n$,

$$\mathbb{P}\left\{ |W_s - W_t| > d_Q\sqrt{\frac{2C_2 v}{n}} + \frac{5C_2\, d_\infty\, v}{6n} \right\} \le 2e^{-v},$$

proving the proposition with $C_3 = 2C_2$. □

B.3 Chaining over a localized class

To control the supremum over a class of score functions, we use the generic chaining result of Dirksen (2015), which is suited to the mixed sub-Gaussian and sub-exponential tail bound in Proposition B.7. We state the form that we need: let γ_1 and γ_2 denote Talagrand's γ-functionals; see, e.g., Dirksen (2015, Section 2). For a metric d, write $\Delta_d(T) = \sup_{s,t\in T} d(s,t)$. Then, for $j \in \{1, 2\}$,

$$\gamma_j(T, d) \le c_j \int_0^{\Delta_d(T)} \big(\log \mathcal{N}(\epsilon, T, d)\big)^{1/j} d\epsilon, \tag{31}$$

by Dirksen (2015, equation (2.3)). We will also use the linear scaling property $\gamma_j(T, \lambda d) = \lambda\gamma_j(T, d)$.

Fact B.8. *Let $(X_t)_{t\in T}$ be a stochastic process on a finite index set with two semimetrics d_1, d_2, and suppose that for all $s, t \in T$ and all $v \ge 0$,*

$$\mathbb{P}\big\{ |X_s - X_t| > \sqrt{v}\, d_2(s,t) + v\, d_1(s,t) \big\} \le 2e^{-v}.$$

There exist universal constants $C, c > 0$ such that, for every $t_0 \in T$ and $v \ge 1$, with probability at least $1 - e^{-v}$,

$$\sup_{t\in T} |X_t - X_{t_0}| \le C\big\{\gamma_2(T, d_2) + \gamma_1(T, d_1)\big\} + c\big\{\sqrt{v}\,\Delta_2(T) + v\,\Delta_1(T)\big\},$$

where $\Delta_j(T)$ is the diameter of T in d_j (see (Dirksen, 2015, Theorem 3.5)).

Recall from Section 4.1 that, for a radius $\rho > 0$, $\mathcal{S}(\rho) = \{s \in \mathcal{S} : \mathcal{E}_K(s) \le \rho^2\}$ is the set of scores whose excess risk is at most ρ^2. Also, define the sup-norm metric entropy

$$\mathcal{H}(\epsilon) = \log \mathcal{N}\big(\epsilon, \mathcal{S}, \|\cdot\|_\infty\big).$$

On a localized class, Proposition 4.1 bounds the d_Q-diameter by a constant multiple of the localization radius. We therefore apply the mixed-tail chaining theorem with this smaller diameter, and use a finite sup-norm net to control the contribution from the finest scales.

Proposition B.9. *Let $0 < \rho \le \sqrt{\kappa}$. There is a constant C_4, depending only on $(M_0, r_{\max})$, such that for every $K \ge 2$, $N \ge 1$, truncation level $\epsilon_0 \in (0, M_0]$, and $v \ge 1$, with probability at least $1 - e^{-v}$,*

$$\sup_{s \in \mathcal{S}(\rho)} |W_s| \le C_4 \Bigg\{ \frac{1}{\sqrt{n}} \int_0^{2\sqrt{C_1}\rho} \sqrt{\mathcal{H}(\epsilon/2)}\, d\epsilon + \frac{1}{n} \int_{\epsilon_0}^{4M_0} \mathcal{H}(\epsilon/4)\, d\epsilon + \frac{\epsilon_0 \mathcal{H}(\epsilon_0/2)}{n} + \sqrt{C_1}\,\rho \sqrt{\frac{v}{n}} + \frac{v}{n} \Bigg\} + 4\epsilon_0. \tag{32}$$

If $\int_0^{4M_0} \mathcal{H}(\epsilon/4)\, d\epsilon < \infty$, then the bound also holds with $\epsilon_0 = 0$, the term $\epsilon_0 \mathcal{H}(\epsilon_0/2)$ and the additive $4\epsilon_0$ being dropped.

Proof. We may assume $\mathcal{H}(\epsilon_0/2) < \infty$ and $\mathcal{S}(\rho) \neq \varnothing$, since otherwise the claim is trivial.

Step 1. Choose an ϵ_0-net $T_0 \subseteq \mathcal{S}(\rho)$ in the sup norm such that $\log |T_0| \le \mathcal{H}(\epsilon_0/2)$. Indeed, an $\epsilon_0/2$-cover of $\mathcal{S}$ can be converted into such an internal ϵ_0-net by choosing a point of $\mathcal{S}(\rho)$ from each covering ball that intersects $\mathcal{S}(\rho)$. Set $T = T_0 \cup \{s^\star\}$; then $\log |T| \le \mathcal{H}(\epsilon_0/2) + \log 2$.

For every $s \in \mathcal{S}(\rho)$, there is $s' \in T_0$ such that $\|s - s'\|_\infty \le \epsilon_0$. By Lemma B.1, $|W_s - W_{s'}| \le 4\epsilon_0$. Since $W_{s^\star} = 0$,

$$\sup_{s \in \mathcal{S}(\rho)} |W_s| \le \max_{t \in T} |W_t - W_{s^\star}| + 4\epsilon_0. \tag{33}$$

Step 2. Define, on T, the semimetrics

$$d_2(s,t) = \frac{C_3}{\sqrt{n}} d_Q(s,t), \qquad d_1(s,t) = \frac{2C_3}{n} \|s - t\|_\infty,$$

with C_3 the constant in Proposition B.7. By Proposition B.7 and Fact B.8, with probability at least $1 - e^{-v}$,

$$\max_{t \in T} |W_t - W_{s^\star}| \le C\big\{\gamma_2(T, d_2) + \gamma_1(T, d_1)\big\} + c\big\{\sqrt{v}\,\Delta_2(T) + v\,\Delta_1(T)\big\}. \tag{34}$$

Step 3. By Proposition 4.1, every $s \in \mathcal{S}(\rho)$ satisfies $d_Q(s, s^\star) \le \sqrt{C_1}\rho$, so the d_Q-diameter of T is at most $2\sqrt{C_1}\rho$ and

$$\Delta_2(T) \le C_3 \cdot \frac{2\sqrt{C_1}\rho}{\sqrt{n}}, \qquad \Delta_1(T) \le C_3 \cdot \frac{4M_0}{n},$$

the latter because $d_\infty(s,t) \le 4M_0$ for all bounded scores. For the γ_2 term, (31) gives

$$\gamma_2(T, d_2) = \frac{C_3}{\sqrt{n}}\, \gamma_2(T, d_Q) \le \frac{C}{\sqrt{n}} \int_0^{2\sqrt{C_1}\rho} \sqrt{\log \mathcal{N}(\epsilon, T, d_Q)}\, d\epsilon.$$

Since $d_Q(s,t) \le \|s - t\|_\infty$, a sup-norm cover is also a d_Q cover. Moreover, an $\epsilon/2$-cover of $\mathcal{S}$ yields an ϵ-cover of $\mathcal{S}(\rho)$, while $s^\star$ requires at most one additional point. Hence

$$\gamma_2(T, d_2) \le \frac{C}{\sqrt{n}} \left\{ \int_0^{2\sqrt{C_1}\rho} \sqrt{\mathcal{H}(\epsilon/2)}\, d\epsilon + 2\sqrt{C_1}\rho \right\}.$$

For the γ_1 term,

$$\log \mathcal{N}(\epsilon, T, 2\|\cdot\|_\infty) = \log \mathcal{N}(\epsilon/2, T, \|\cdot\|_\infty) \le \min\{\mathcal{H}(\epsilon_0/2), \mathcal{H}(\epsilon/4)\} + \log 2.$$

The diameter of T in the metric $2\|\cdot\|_\infty$ is at most $4M_0$. Consequently,

$$\gamma_1(T, d_1) \le \frac{C_3 c_1}{n} \int_0^{4M_0} \log \mathcal{N}(\epsilon, T, 2\|\cdot\|_\infty) d\epsilon \le \frac{C}{n} \left\{ \epsilon_0 \mathcal{H}(\epsilon_0/2) + \int_{\epsilon_0}^{4M_0} \mathcal{H}(\epsilon/4) d\epsilon + 4M_0 \log 2 \right\}.$$

Step 4. Substituting the bounds from Step 3 into (34) and then into (33), we obtain (32) up to the additional terms $C\rho/\sqrt{n}$ and CM_0/n. Since $v \ge 1$ and $\sqrt{C_1} \ge 1$, these terms can be absorbed into $\sqrt{C_1}\rho\sqrt{v/n}$ and v/n, respectively, after enlarging the constant. This proves (32) with a constant C_4 depending only on $(M_0, r_{\max})$.

Step 5. Assume that $\int_0^{4M_0} \mathcal{H}(\epsilon/4) d\epsilon < \infty$. Since $\mathcal{H}$ is nonincreasing,

$$\epsilon \mathcal{H}(\epsilon/2) \le 2 \int_{\epsilon/2}^{\epsilon} \mathcal{H}(\epsilon'/2) d\epsilon' \longrightarrow 0 \qquad \text{as } \epsilon \downarrow 0.$$

Hence, as $\epsilon_0 \downarrow 0$, the right-hand side of (32) converges to the same expression with $\epsilon_0 = 0$: the entropy integral extends to zero, while $\epsilon_0 \mathcal{H}(\epsilon_0/2)/n$ and $4\epsilon_0$ vanish.

Therefore, for any $\zeta > 0$, we may choose a deterministic $\epsilon_0 > 0$ small enough that the right-hand side of (32) is at most the bound with $\epsilon_0 = 0$ plus ζ. Since (32) holds with probability at least $1 - e^{-v}$ for this choice of ϵ_0, the same probability bound holds with an additional ζ. Letting $\zeta \downarrow 0$ gives the claim for $\epsilon_0 = 0$. □

B.4 Critical radius and proof of Theorem 4.2

We next incorporate the localized bound into the peeling argument. The following definition assigns confidence levels to the dyadic radii.

Definition B.10. *Fix a confidence level $\eta \in (0,1)$ and recall $\kappa = 4M_0$. For $0 < \delta \le \sqrt{\kappa}$, let*

$$J_\delta = \left\lceil \log_2 \frac{\sqrt{\kappa}}{\delta} \right\rceil, \qquad t_\delta = 1 + \log \frac{J_\delta + 1}{\eta}.$$

We call δ an admissible radius if there is a truncation level $\epsilon_0 \in [0, M_0]$ such that

$$C_4 \left\{ \frac{1}{\sqrt{n}} \int_0^{2\sqrt{C_1}\,\delta} \sqrt{\mathcal{H}(\epsilon/2)} d\epsilon + \frac{1}{n} \int_{\epsilon_0}^{4M_0} \mathcal{H}(\epsilon/4) d\epsilon + \frac{\epsilon_0 \mathcal{H}(\epsilon_0/2)}{n} + \sqrt{C_1}\,\delta \sqrt{\frac{t_\delta}{n}} + \frac{t_\delta}{n} \right\} + 4\epsilon_0 \le \frac{\delta^2}{16}, \tag{35}$$

where C_1 and C_4 are constants in Propositions 4.1 and B.9 respectively. The choice $\epsilon_0 = 0$ is allowed when $\int_0^{4M_0} \mathcal{H}(\epsilon/4) d\epsilon < \infty$; in that case the term $\epsilon_0 \mathcal{H}(\epsilon_0/2)$ is defined to be zero. Write

$$\mathcal{A}_{n,\eta} = \left\{ \delta \in (0, \sqrt{\kappa}] : \delta \text{ is admissible} \right\}.$$

The envelope and critical radius. We now make the high-probability bound $\mathfrak{R}_n(\rho)$ from Section 4.1 explicit. Fix $\eta \in (0,1)$, and let $F_n(\rho, \epsilon_0)$ denote the left-hand side of (35), with $\delta = \rho$ and $t_\delta = t_\rho$. Recall from (6) that

$$\Delta_n(\rho) = \sup_{s \in \mathcal{S}(\rho)} |(\widehat{L}_K - L_K)(s) - (\widehat{L}_K - L_K)(s^\star)|,$$

which equals $\sup_{s \in \mathcal{S}(\rho)} |W_s|$ by the definition of W_s. For each fixed ϵ_0, Proposition B.9 bounds $\Delta_n(\rho)$ by $F_n(\rho, \epsilon_0)$ with high probability, while Lemma B.1 gives the deterministic bound $\Delta_n(\rho) \le 2\kappa$. We therefore define

$$\mathfrak{R}_n(\rho) = \begin{cases} \min\{2\kappa,\ 2 \inf_{\epsilon_0} F_n(\rho, \epsilon_0)\}, & 0 < \rho \le \sqrt{\kappa}, \\ 2\kappa, & \rho > \sqrt{\kappa}, \end{cases} \tag{36}$$

where the infimum is over the truncation levels allowed in Definition B.10. The factor 2 allows us to choose a truncation level arbitrarily close to the infimum without assuming that the infimum is attained.

The critical radius from Section 4.1 is therefore given by

$$\delta_n = \inf\{\rho > 0 : \mathfrak{R}_n(\rho) \le \rho^2/16\}. \tag{37}$$

Moreover, if $\rho \le \sqrt{\kappa}$ satisfies $\mathfrak{R}_n(\rho) \le \rho^2/16$, then the branch 2κ cannot be active. Hence $2\inf_{\epsilon_0} F_n(\rho, \epsilon_0) \le \rho^2/16$, and we can choose ϵ_0 such that $F_n(\rho, \epsilon_0) \le \rho^2/16$. Thus every such ρ is admissible in the sense of Definition B.10.

Lemma B.11. *For each $\delta \in \mathcal{A}_{n,\eta}$, with probability at least $1-\eta$,*

$$|W_s| \le \tfrac{1}{4}\mathcal{E}_K(s) + \delta^2/16 \qquad \textit{for every } s \in \mathcal{S}.$$

The same conclusion holds with δ_n from (37) in place of δ.

We apply the localized bound at geometrically increasing radii to prove Lemma B.11.

Proof. Let $\delta \in \mathcal{A}_{n,\eta}$ satisfy (35) with a deterministic truncation level ϵ_0, and let J_δ, t_δ be as in Definition B.10. Recall from the discussion after Lemma B.1 that $\mathcal{S} = \mathcal{S}(\sqrt{\kappa})$.

Define radii $\rho_0 = \delta$ and $\rho_j = (2^j\delta) \wedge \sqrt{\kappa}$ for $1 \le j \le J_\delta$; by the definition of J_δ, $\rho_{J_\delta} = \sqrt{\kappa}$. Apply Proposition B.9 with radius ρ_j, truncation ϵ_0, and $v = t_\delta \ge 1$, for $j = 0, \dots, J_\delta$. Each application fails with probability at most $e^{-t_\delta} = \eta e^{-1}/(J_\delta + 1)$, so all $J_\delta + 1$ hold simultaneously with probability at least $1 - \eta/e \ge 1 - \eta$. Assume they all hold.

Write the right-hand side of (32) at radius ρ and $v = t_\delta$ as $G(\rho) + E$, where

$$G(\rho) = C_4 \left\{ \frac{1}{\sqrt{n}} \int_0^{2\sqrt{C_1}\rho} \sqrt{\mathcal{H}(\epsilon/2)}d\epsilon + \sqrt{C_1}\,\rho\sqrt{\frac{t_\delta}{n}} \right\}$$

collects the two terms that depend on ρ, while E contains the remaining terms, which are independent of ρ. Since $\sqrt{\mathcal{H}(\cdot/2)}$ is nonincreasing, the average $\rho^{-1}\int_0^{2\sqrt{C_1}\rho} \sqrt{\mathcal{H}(\epsilon/2)}d\epsilon$ is nonincreasing in ρ, so $G(\rho)/\rho$ is nonincreasing. Admissibility (35) says exactly $G(\delta) + E \le \delta^2/16$; in particular $E \le \delta^2/16$, and for $\rho \ge \delta$,

$$G(\rho) + E \le \frac{\rho}{\delta}\,G(\delta) + E \le \frac{\rho}{\delta}\Big(\frac{\delta^2}{16} - E\Big) + E \le \frac{\rho\,\delta}{16}.$$

Now take any $s \in \mathcal{S}$. If $\mathcal{E}_K(s) \le \delta^2$, then $s \in \mathcal{S}(\rho_0)$ and $|W_s| \le G(\delta) + E \le \delta^2/16$. Otherwise there is a unique $1 \le j \le J_\delta$ with $\rho_{j-1}^2 < \mathcal{E}_K(s) \le \rho_j^2$, so $s \in \mathcal{S}(\rho_j)$, and since $\rho_j \le 2\rho_{j-1}$ in all cases,

$$|W_s| \le G(\rho_j) + E \le \frac{\rho_j\delta}{16} \le \frac{\rho_j^2}{16} \le \frac{4\,\rho_{j-1}^2}{16} < \frac{1}{4}\,\mathcal{E}_K(s).$$

Thus, for each deterministic $\delta \in \mathcal{A}_{n,\eta}$, with probability at least $1-\eta$,

$$|W_s| \le \frac{1}{4}\mathcal{E}_K(s) + \frac{\delta^2}{16} \qquad \text{for every } s \in \mathcal{S}.$$

Because δ_n is defined as an infimum, it need not itself belong to the fixed-point set in (37) and hence need not be admissible. Suppose first that the fixed-point set contains a radius at most $\sqrt{\kappa}$. Choose a deterministic sequence of fixed-point radii $\delta_j \le \sqrt{\kappa}$ such that $\delta_j \downarrow \delta_n$. Each δ_j is admissible by the discussion following (37), so the bound just established gives, with probability at least $1-\eta$,

$$\sup_{s\in\mathcal{S}} \left[|W_s| - \frac{1}{4}\mathcal{E}_K(s) \right] \le \frac{\delta_j^2}{16}.$$

As $\delta_j \downarrow \delta_n$, these events decrease to

$$\left\{ \sup_{s\in\mathcal{S}} \left[|W_s| - \frac{1}{4}\mathcal{E}_K(s) \right] \le \frac{\delta_n^2}{16} \right\},$$

which therefore also has probability at least $1-\eta$.

If the fixed-point set contains no radius at most $\sqrt{\kappa}$, then (36) gives $\delta_n = \sqrt{32\kappa}$. In this case, $|W_s| \le 2\kappa = \delta_n^2/16$ for every $s \in \mathcal{S}$, so the same bound holds deterministically. This completes the proof. □

We now complete the proof of Theorem 4.2. The uniform deviation bound and empirical optimality yield an excess-risk bound for the empirical InfoNCE minimizer, which Proposition 4.1 converts into a PMI error bound.

Proof of Theorem 4.2. Work on the event of Lemma B.11, of probability at least $1-\eta$. For every $s \in \mathcal{S}$, the optimality of $\widehat{s}$ gives $\widehat{L}_K(\widehat{s}) \le \widehat{L}_K(s)$, that is,

$$\mathcal{E}_K(\widehat{s}) \le \mathcal{E}_K(s) + W_s - W_{\widehat{s}} \le \mathcal{E}_K(s) + \frac{1}{4}\mathcal{E}_K(s) + \frac{1}{4}\mathcal{E}_K(\widehat{s}) + \frac{\delta_n^2}{8},$$

whence $\mathcal{E}_K(\widehat{s}) \le \frac{5}{3}\mathcal{E}_K(s) + \frac{\delta_n^2}{6}$; taking the infimum over $s \in \mathcal{S}$ proves the first inequality, and Proposition 4.1 converts it into the second. If $s^\star + c \in \mathcal{S}$, choose $s = s^\star + c$: then $\mathcal{E}_K(s) = 0$ and $W_s = 0$, so $\mathcal{E}_K(\widehat{s}) \le \frac{1}{4}\mathcal{E}_K(\widehat{s}) + \delta_n^2/16$, giving $\mathcal{E}_K(\widehat{s}) \le \delta_n^2/12$, and Proposition 4.1 then gives $\mathrm{Err}_Q(\widehat{s}) \le C\delta_n^2$. □

C ReLU NETWORKS AND MINIMAX RATES

In this section, we provide the proofs of Corollaries 4.3 and 4.4 and Theorems 4.5 and 4.6.

C.1 ReLU SCORE CLASSES

This subsection specializes the general oracle inequality to the neural-network score classes in Section 4.2. We first derive an explicit critical radius for a class with logarithmic covering entropy, and then combine it with covering number bounds for ReLU networks.

Lemma C.1. *Suppose that for some $V \ge 1$ and $A \ge e$,*

$$\mathcal{H}(\epsilon) \le V \log(A/\epsilon), \qquad 0 < \epsilon \le 1.$$

Then there is a constant C, depending only on $(M_0, r_{\min}, r_{\max})$, such that, for every $\eta \in (0,1)$, $K \ge 2$, $N \ge 1$, and $n = NK \ge 2$, with probability at least $1-\eta$,

$$\mathcal{E}_K(\widehat{s}) \le C\left\{\inf_{s\in\mathcal{S}} \mathcal{E}_K(s) + \frac{V\log(An) + \log(1/\eta)}{n}\right\},$$

and the same right-hand side, with a possibly larger C, bounds $\mathrm{Err}_Q(\widehat{s})$.

We use the following elementary entropy integrals. For $A' \ge e$ and $0 < D \le A'$,

$$\int_0^D \sqrt{\log\frac{A'}{\epsilon}}\,d\epsilon \le D\left(\sqrt{\log\frac{A'}{D}} + 1\right), \tag{38}$$

which follows after writing $\epsilon = Du$ and using $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ and $\int_0^1 \sqrt{\log(1/u)}du \le 1$. Also,

$$\int_0^D \log\frac{A'}{\epsilon}\,d\epsilon = D\left(\log\frac{A'}{D} + 1\right). \tag{39}$$

Proof. By monotonicity of covering numbers, $\mathcal{H}(\epsilon) \le \mathcal{H}(1) \le V\log A$ for $\epsilon > 1$. Put

$$R_n = \frac{V\log(An) + \log(1/\eta)}{n}, \qquad \delta^2 = C_5 R_n,$$

where $C_5 = C_5(M_0, r_{\min}, r_{\max})$ will be chosen below. We first handle the case $\delta^2 > \kappa$. Every score in the bounded class satisfies $\mathcal{E}_K(s) \le \kappa$ and $\mathrm{Err}_Q(s) \le \|s - s^\star\|_\infty^2 \le 4M_0^2 = \frac{\kappa^2}{4}$. Since

$R_n = \delta^2/C_5 > \kappa/C_5$, both quantities are bounded by a constant, depending only on M_0 through $\kappa = 4M_0$, times R_n. The claimed result follows after enlarging its constant. We may therefore assume that $\delta \le \sqrt{\kappa}$ and verify (35) with $\epsilon_0 = 0$.

The square-root entropy integral. Let $D = (2\sqrt{C_1}\,\delta) \wedge 2$. For $0 < \epsilon \le D$, the entropy assumption and (38), with $A' = 2A$, give

$$\int_0^D \sqrt{\mathcal{H}(\epsilon/2)}d\epsilon \le \sqrt{V}\int_0^D \sqrt{\log(2A/\epsilon)}d\epsilon \le D\sqrt{V}\left\{\sqrt{\log(2A/D)}+1\right\}.$$

Because $V \ge 1$, $C_5 \ge 1$, and $A \ge e$, the definition of δ implies $\delta^2 \ge C_5/n \ge 1/n$, hence $D \ge 2n^{-1/2}$ whenever $D < 2$; in that case $\log(2A/D) \le \log(A\sqrt{n}) \le \log(An)$, while for $D = 2$ we have $\log(2A/D) = \log A \le \log(An)$ directly. This implies $\int_0^D \sqrt{\mathcal{H}(\epsilon/2)}d\epsilon \le C\sqrt{C_1}\,\delta\sqrt{V\log(An)}$. If $2\sqrt{C_1}\delta > 2$, the part of the original integral from 2 to $2\sqrt{C_1}\delta$ is bounded, using $\mathcal{H}(\epsilon/2) \le V\log A$, by $(2\sqrt{C_1}\delta - 2)\sqrt{V\log A} \le 2\sqrt{C_1}\delta\sqrt{V\log(An)}$. Combining the two ranges, for large enough C_5,

$$\frac{C_4}{\sqrt{n}}\int_0^{2\sqrt{C_1}\delta}\sqrt{\mathcal{H}(\epsilon/2)}d\epsilon \le CC_4\sqrt{C_1}\,\delta\sqrt{R_n} = \frac{CC_4\sqrt{C_1}}{\sqrt{C_5}}\,\delta^2 \le \delta^2/80.$$

The first-power entropy integral. For $0 < \epsilon \le 4$, the entropy assumption and (39) give $\int_0^4 \mathcal{H}(\epsilon/4)d\epsilon \le 4V(\log A + 1)$. For $4 < \epsilon \le 4M_0$, monotonicity gives $\mathcal{H}(\epsilon/4) \le \mathcal{H}(1) \le V\log A$. Hence

$$\int_0^{4M_0}\mathcal{H}(\epsilon/4)d\epsilon \le C(1+M_0)V\log A.$$

After multiplication by C_4/n, this is at most $CC_4(1+M_0)R_n = CC_4(1+M_0)\delta^2/C_5 \le \delta^2/80$.

The probability terms. The bound $\delta^2 \ge C_5/n$ implies $\sqrt{\kappa}/\delta \le \sqrt{\kappa n/C_5}$, and hence

$$J_\delta \le 1+\log_2(\kappa n), \qquad t_\delta \le C\{1+\log(1/\eta)+\log\log(en)\} \le C\{V\log(An)+\log(1/\eta)\} = CnR_n.$$

It follows that $C_4\sqrt{C_1}\,\delta\sqrt{\frac{t_\delta}{n}} \le \frac{CC_4\sqrt{C_1}}{\sqrt{C_5}}\,\delta^2$ and $\frac{C_4 t_\delta}{n} \le \frac{CC_4}{C_5}\,\delta^2$. For sufficiently large C_5, both terms are at most $\delta^2/80$. Combining these bounds gives (35), so δ is an admissible radius.

Lemma B.11, together with the same empirical-optimality calculation used in the proof of Theorem 4.2, gives

$$\mathcal{E}_K(\widehat{s}) \le \frac{5}{3}\inf_{s\in\mathcal{S}}\mathcal{E}_K(s) + \frac{C_5R_n}{6},$$

and Proposition 4.1 gives the same form for $\mathrm{Err}_Q(\widehat{s})$, with a possibly larger constant. This proves the result. □

We next give the network covering bounds used below. Throughout this section, depth counts affine layers, including the final output layer; ReLU is applied only at the hidden layers.

Lemma C.2. *Let $B \ge 1$, $W \ge d$, $S \ge \max\{W, L\}$, and put $D = \min\{S, W^2L\}$. There is a universal constant C such that, for $0 < \epsilon < 1/2$,*

$$\log\mathcal{N}\big(\epsilon, \mathcal{R}(d,W,L,B,S), \|\cdot\|_\infty\big) \le CD\log\left(\frac{(W+1)^L B^L}{\epsilon}\right). \tag{40}$$

Let $\mathcal{R}_p(d,W,L,B,S)$ be the corresponding class with output in $\mathbb{R}^p$, where $W \ge \max\{d,p\}$ and S counts the nonzero parameters of the whole network. Then

$$\log\mathcal{N}\big(\epsilon, \mathcal{R}_p(d,W,L,B,S), \sup_x\|\cdot\|_2\big) \le CD\log\left(\frac{\sqrt{p}\,(W+1)^L B^L}{\epsilon}\right). \tag{41}$$

Any subclass, for example the subclass satisfying $\sup_x\|f(x)\|_2 \le R$, obeys the same bound.

Proof. Equation (40) follows from the L^∞ case of Ou & Bölcskei (2026, Theorem 5.1). Their network class imposes the same restrictions on depth, width, parameter magnitude, and the total number of nonzero weights and biases. Taking the norm exponent in their theorem to be infinity gives

$$\log \mathcal{N}\big(\epsilon, \mathcal{R}(d, W, L, B, S), \|\cdot\|_\infty\big) \le CD \log\left(\frac{(W+1)^L B^L}{\epsilon}\right),$$

where $D = \min\{S, W^2 L\}$.

For the vector-valued class, fix two networks $f, \widetilde{f} \in \mathcal{R}_p(d, W, L, B, S)$ with the same architecture, and suppose that every corresponding parameter differs by at most q. For each output coordinate $j = 1, \dots, p$, the parameter perturbation bound of Ou & Bölcskei (2026, Lemma 2.2) gives

$$\sup_x |f_j(x) - \widetilde{f}_j(x)| \le L(W+1)^L B^{L-1} q.$$

Hence

$$\sup_x \|f(x) - \widetilde{f}(x)\|_2 \le \sqrt{p}\, L(W+1)^L B^{L-1} q.$$

Choose

$$q = \frac{\epsilon}{\sqrt{p}\, L(W+1)^L B^{L-1}}.$$

Then rounding each nonzero parameter toward zero on a grid of mesh at most q produces an ϵ-cover in the metric $\sup_x \|\cdot\|_2$.

Since $d, p \le W$, each affine layer contains at most a constant multiple of W^2 parameter locations, including biases. Thus the whole network contains at most $CW^2 L$ possible parameter locations. At most S of these are nonzero, so the number of active parameter locations is at most a constant multiple of

$$D = \min\{S, W^2 L\}.$$

The number of possible supports and quantized parameter values can therefore be bounded by the same counting argument as in the proof of Ou & Bölcskei (2026, Theorem 5.1). Since each nonzero parameter belongs to $[-B, B]$, the number of grid values for one active parameter is at most

$$1 + \frac{2B}{q} \le C \frac{\sqrt{p}\, L(W+1)^L B^L}{\epsilon}.$$

Consequently,

$$\log \mathcal{N}\big(\epsilon, \mathcal{R}_p(d, W, L, B, S), \sup_x \|\cdot\|_2\big) \le CD \log\left(\frac{C\sqrt{p}\, L(W+1)^L B^L}{\epsilon}\right).$$

Because $W \ge 1$ and $L \ge 1$, the factors C and L can be absorbed into the logarithm after increasing the universal constant. This gives

$$\log \mathcal{N}\big(\epsilon, \mathcal{R}_p(d, W, L, B, S), \sup_x \|\cdot\|_2\big) \le CD \log\left(\frac{\sqrt{p}\,(W+1)^L B^L}{\epsilon}\right),$$

which is (41).

Finally, for a nonempty subclass $\mathcal{A} \subseteq \mathcal{R}_p$, cover $\mathcal{R}_p$ at radius $\epsilon/2$ and choose a point of $\mathcal{A}$ from each covering ball that meets $\mathcal{A}$. Then it follows that $\mathcal{N}(\epsilon, \mathcal{A}, d) \le \mathcal{N}(\epsilon/2, \mathcal{R}_p, d)$. Thus the entropy bound in the lemma holds for subclasses after increasing the universal constant. □

The result for the class of joint-input ReLU networks follows directly from Lemmas C.1 and C.2.

Proof of Corollary 4.3. Since $\mathcal{S}_{\text{net}}$ is a subclass of the scalar ReLU class, Lemma C.2 gives $\mathcal{H}(\epsilon) \le CD \log(A_0/\epsilon)$ for $0 < \epsilon < 1/2$, where $D = \min\{S, W^2 L\}$ and $A_0 = (W+1)^L B^L$. For $1/2 \le \epsilon \le 1$, the same bound, with a larger universal constant, follows from monotonicity of covering numbers. Thus Lemma C.1, with $V = CD$ and $A = eA_0$, gives

$$\mathcal{E}_K(\widehat{s}),\ \mathrm{Err}_Q(\widehat{s}) \le C \left\{ \inf_{s \in \mathcal{S}_{\text{net}}} \mathcal{E}_K(s) + \frac{D \log\big(en(W+1)^L B^L\big) + \log(1/\eta)}{n} \right\}.$$

By (14),

$$\inf_{s\in\mathcal{S}_{\rm net}} \mathcal{E}_K(s) \le \frac{1}{2} \inf_{s\in\mathcal{S}_{\rm net},\, c\in\mathbb{R}} \|s - s^\star - c\|_\infty^2 = \frac{1}{2}\epsilon_{\rm net}^2.$$

Substituting this bound proves the corollary. □

The result for the CLIP-style score class follows by transferring covers of the two encoders to a cover of their inner-product scores.

Proof of Corollary 4.4. Write $D = \min\{S, W^2L\}$ and $A_0 = (W+1)^L B^L$. For each encoder, writing v_ℓ for the output of layer ℓ (with $v_0 = x$), the layerwise bound $\|v_\ell\|_\infty \le WB\|v_{\ell-1}\|_\infty + B$, with $\|v_0\|_\infty \le 1$, gives $\sup_x \|f(x)\|_\infty \le A_0$, and the same bound holds for g.

Suppose that corresponding parameters in two networks of the same architecture differ by at most q. Applying the parameter perturbation bound of Ou & Bölcskei (2026, Lemma 2.2) to each output coordinate gives $\sup_x \|f(x) - f'(x)\|_\infty \le LA_0 q/B$, and similarly for g. Hence

$$\|s_{f,g} - s_{f',g'}\|_\infty \le \frac{2pLA_0^2}{B\tau}\, q. \tag{42}$$

For fixed encoder depths, the two networks have at most $4W^2L$ parameter locations and at most $4D$ nonzero parameters. There are at most L^2 choices of depths, and widths can be padded to W using zero parameters. Since $D \ge \max\{W, L\}$, the support and depth choices contribute at most $CD\log(C(D+1))$ to the logarithm of the cover size.

For $0 < \epsilon \le 1$, round every nonzero parameter toward zero on a grid with spacing $q = B\min\{1, \epsilon\tau/(8pLA_0^2)\}$. This preserves the parameter bound and sparsity, while (42) gives score error at most $\epsilon/4$. Each active parameter has at most $3 + 16pLA_0^2/(\epsilon\tau)$ possible grid values. Therefore

$$\log \mathcal{N}\big(\epsilon, \mathcal{S}_{\rm IP}, \|\cdot\|_\infty\big) \le CD\log\left(\frac{C(D+1)^2(1+pA_0^2/\tau)}{\epsilon}\right),$$

after increasing the universal constant. Here the grid scores need not belong to $\mathcal{S}_{\rm IP}$: from each $\epsilon/4$-ball that meets $\mathcal{S}_{\rm IP}$, choose one member of $\mathcal{S}_{\rm IP}$. The triangle inequality gives an internal $\epsilon/2$-cover.

Since all scores in $\mathcal{S}_{\rm IP}$ and $s^\star$ are bounded by M_0, Lemma C.1, with $V = CD$ and $A = C(D+1)^2(1+pA_0^2/\tau)$, gives

$$\mathrm{Err}_Q(\widehat{s}) \le C\left\{\inf_{s\in\mathcal{S}_{\rm IP}} \mathcal{E}_K(s) + \frac{D\log\big(Cn(D+1)^2(1+pA_0^2/\tau)\big) + \log(1/\eta)}{n}\right\}.$$

If $D \le n$, the logarithm is bounded, up to a universal constant, by $\log\big(en(1+pA_0^2/\tau)\big)$. If $D > n$, the desired bound follows from $\mathrm{Err}_Q(\widehat{s}) \le 4M_0^2$ and $D\log(en(1+pA_0^2/\tau))/n \ge 1$.

Finally, (14) gives

$$\inf_{s\in\mathcal{S}_{\rm IP}} \mathcal{E}_K(s) \le \frac{1}{2} \inf_{s\in\mathcal{S}_{\rm IP},\, c\in\mathbb{R}} \|s - s^\star - c\|_\infty^2 = \frac{1}{2}\epsilon_{\rm IP}^2.$$

Substituting $A_0^2 = (W+1)^{2L}B^{2L}$ proves the corollary. □

C.2 Hölder approximation and proof of Theorem 4.5

We work with the Hölder class from Section 4.2, with $r = \lceil\alpha\rceil - 1$ and $\beta = \alpha - r \in (0, 1]$. Throughout this subsection, constants may depend on α, d_X, d_Y, M and, in the two-encoder case, on the fixed temperature $\tau > 0$. We first construct the required ReLU approximations and specify their network sizes, and then apply Corollaries 4.3 and 4.4.

Lemma C.3. *For every fixed integer $q \ge 1$ and $0 < \epsilon < 1/2$, the function $(v_1, \dots, v_q) \mapsto \prod_{j=1}^q v_j$ on $[-1, 1]^q$ can be approximated uniformly to error at most ϵ by a ReLU network with width at most C_q, depth and number of nonzero parameters at most $C_q \log(1/\epsilon)$, and all parameters bounded by C_q.*

Proof. The case $q = 1$ is the identity map. Suppose $q \geq 2$. By Schmidt-Hieber (2020, Lemma A.2), for every integer $m \geq 1$ there is a ReLU network M_m on $[0,1]^2$ such that

$$\sup_{a,b\in[0,1]} |M_m(a,b) - ab| \leq 2^{-m},$$

with constant width, depth and number of nonzero parameters of order $O(m)$, and uniformly bounded parameters.

For $u, v \in [-2,2]$, write $u = u_+ - u_-$ and $v = v_+ - v_-$, where $u_\pm, v_\pm \in [0,2]$. Define

$$\widetilde{M}_m(u,v) = 4\Big[M_m(u_+/2, v_+/2) - M_m(u_+/2, v_-/2) - M_m(u_-/2, v_+/2) + M_m(u_-/2, v_-/2)\Big].$$

Since

$$uv = 4\Big[(u_+/2)(v_+/2) - (u_+/2)(v_-/2) - (u_-/2)(v_+/2) + (u_-/2)(v_-/2)\Big],$$

we have

$$\sup_{|u|,|v|\leq 2} |\widetilde{M}_m(u,v) - uv| \leq 16 \cdot 2^{-m}.$$

The maps $u \mapsto u_+$ and $u \mapsto u_-$ are represented exactly by ReLU units, so $\widetilde{M}_m$ has constant width, depth and sparsity $O(m)$, and uniformly bounded parameters.

Set $\epsilon' = \epsilon/q$ and choose $m = \lceil \log_2(16/\epsilon') \rceil$. Then $\widetilde{M}_m$ approximates multiplication on $[-2,2]^2$ with error at most ϵ'. Starting with $\widetilde{P}_1 = v_1$, define recursively $\widetilde{P}_j = \widetilde{M}_m(\widetilde{P}_{j-1}, v_j)$ for $j = 2, \ldots, q$. If $P_j = \prod_{i=1}^{j} v_i$, then

$$|\widetilde{P}_j - P_j| \leq \epsilon' + |\widetilde{P}_{j-1} - P_{j-1}|,$$

because $|v_j| \leq 1$. Hence $|\widetilde{P}_j - P_j| \leq (j-1)\epsilon' < \epsilon$. In particular, $|\widetilde{P}_j| \leq 1 + \epsilon < 2$, so every multiplication network is evaluated within $[-2,2]^2$.

The unused inputs and the signed intermediate values can be passed through the network using their positive and negative parts. Since q is fixed, this changes the width and parameter count only by a constant depending on q. Finally, $m \leq C_q \log(1/\epsilon)$ for $0 < \epsilon < 1/2$, which gives the stated bounds on width, depth, sparsity, and parameter magnitude. □

We next combine these product networks with local Taylor polynomials.

Lemma C.4. *For every integer $m \geq 2$ and every h with $\|h\|_{C^\alpha} \leq M$, there exist a joint-input ReLU network s_m and two ReLU encoders f_m, g_m with p outputs such that*

$$\|s_m - h\|_\infty \leq Cm^{-\alpha}, \qquad \left\| \frac{\langle f_m, g_m \rangle}{\tau} - h \right\|_\infty \leq Cm^{-\alpha}, \tag{43}$$

where the networks can be chosen with

$$L \leq C \log m, \qquad W \leq Cm^{d_X + d_Y}, \qquad S \leq Cm^{d_X+d_Y} \log m, \qquad B \leq Cm, \tag{44}$$

and the encoders with

$$p \leq Cm^{\min\{d_X, d_Y\}}.$$

The bounds on L, W, S, B apply to each encoder separately and are uniform over all h with $\|h\|_{C^\alpha} \leq M$.

Proof. Write $d = d_X + d_Y$. For the joint-input approximation, apply Schmidt-Hieber (2020, Theorem 5). Let $\nu_m = \lceil c_0 m^d \rceil$ and choose $k_m = \lceil (d+\alpha) \log_2 m \rceil + c_1$, where c_0 and c_1 depend only on the fixed parameters and are large enough to satisfy the conditions of the theorem for every $m \geq 2$. Its approximation bound gives

$$\|s_m - h\|_\infty \leq C\{\nu_m 2^{-k_m} + \nu_m^{-\alpha/d}\} \leq Cm^{-\alpha}.$$

The same theorem gives depth $O(\log m)$, width $O(m^d)$, sparsity $O(m^d \log m)$, and uniformly bounded parameters. Thus the joint-input network satisfies the bounds in (44).

For the two-encoder approximation, let $r = \lceil \alpha \rceil - 1$ and define $\lambda_i(t) = (1 - m|t - i/m|)_+$ for $i = 0, \dots, m$. For grid indices $a \in \{0, \dots, m\}^{d_X}$ and $b \in \{0, \dots, m\}^{d_Y}$, let $\phi_a(x) = \prod_{j=1}^{d_X} \lambda_{a_j}(x_j)$ and $\phi_b(y) = \prod_{j=1}^{d_Y} \lambda_{b_j}(y_j)$. For multi-indices u, v with $|u| + |v| \le r$, set $c_{a,b,u,v} = \partial_x^u \partial_y^v h(a/m, b/m)/(u!v!)$, and set $c_{a,b,u,v} = 0$ when $|u| + |v| > r$. Define

$$H_m(x, y) = \sum_{a,b} \phi_a(x)\phi_b(y) \sum_{|u|+|v| \le r} c_{a,b,u,v}(x - a/m)^u (y - b/m)^v.$$

The functions $\phi_a(x)\phi_b(y)$ are nonnegative and sum to one. Whenever one of them is nonzero, (x, y) is within distance $1/m$ of $(a/m, b/m)$ in the maximum norm. The Hölder condition and Taylor's formula therefore give $\|H_m - h\|_\infty \le Cm^{-\alpha}$.

Suppose first that $d_X \le d_Y$. Define $A_{a,u}(x) = \phi_a(x)(x - a/m)^u$ and $B_{b,v}(y) = \phi_b(y)(y - b/m)^v$. There are $O(m^{d_X})$ functions $A_{a,u}$ and $O(m^{d_Y})$ functions $B_{b,v}$, and each is a product of a fixed number of the functions λ_i and coordinate differences. Choose

$$\gamma = c_2 m^{-(\alpha+d)},$$

where $c_2 > 0$ is small enough that $\gamma < 1/2$ for every $m \ge 2$. Lemma C.3 gives ReLU approximations $\widetilde{A}_{a,u}$ and $\widetilde{B}_{b,v}$ with uniform error at most γ, depth and sparsity $O(\log m)$, constant width, and parameters bounded by Cm.

Define

$$f_m(x) = \big(\widetilde{A}_{a,u}(x)\big)_{a,u}, \qquad g_m(y) = \left(\tau \sum_{b,v} c_{a,b,u,v} \widetilde{B}_{b,v}(y)\right)_{a,u}.$$

The common output dimension is at most Cm^{d_X}. Since $|A_{a,u}|, |B_{b,v}| \le 1$ and $\gamma < 1$,

$$|\widetilde{A}_{a,u}\widetilde{B}_{b,v} - A_{a,u}B_{b,v}| \le 3\gamma.$$

There are at most Cm^d terms and $|c_{a,b,u,v}| \le M$, so

$$\left\| \frac{\langle f_m, g_m \rangle}{\tau} - H_m \right\|_\infty \le CMm^d\gamma \le Cm^{-\alpha}.$$

Together with $\|H_m - h\|_\infty \le Cm^{-\alpha}$, this proves the two-encoder approximation bound.

Each $\widetilde{B}_{b,v}$ is computed once, and a single final affine map produces all coordinates of g_m. The feature networks use $O((m^{d_X} + m^{d_Y}) \log m)$ nonzero parameters, while the final map uses $O(m^d)$ parameters. Hence each encoder has depth at most $C \log m$, width at most $C(m^{d_X} + m^{d_Y}) \le Cm^d$, and sparsity at most $Cm^d \log m$. The functions λ_i have parameters of order m, the product networks have uniformly bounded parameters, and the final affine coefficients are bounded by τM, so $B \le Cm$. Signed features can be passed through hidden layers using their positive and negative parts, changing these bounds only by constant factors.

If $d_Y < d_X$, interchange the roles of the two modalities. Thus $p \le Cm^{\min\{d_X, d_Y\}}$, and all the bounds in (44) hold uniformly over $\|h\|_{C^\alpha} \le M$. □

Now, we proceed with the proof of Theorem 4.5.

Proof of Theorem 4.5. Write $d = d_X + d_Y$. For sufficiently large m, Lemma C.4 gives approximating scores with uniform error at most one. Since $\|s^\star\|_\infty \le M$, both approximating scores are bounded by $M + 1 = M_0$ and hence belong to the score classes in Section 4.2, after enlarging the constants in Lemma C.4. Moreover, $\epsilon_{\text{net}}, \epsilon_{\text{IP}} \le Cm^{-\alpha}$.

For the network sizes in Lemma C.4, $D = \min\{S, W^2L\} \le S \le Cm^d \log m$, and the logarithmic factors in Corollaries 4.3 and 4.4 are both bounded by $C\{\log(en) + (\log m)^2\}$, where C may depend on τ for the CLIP-style class. Thus either estimator satisfies, with probability at least $1 - \eta$,

$$\mathrm{Err}_Q(\widehat{s}) \le C\left\{ m^{-2\alpha} + \frac{m^d \log m \,\{\log(en) + (\log m)^2\}}{n} + \frac{\log(1/\eta)}{n} \right\}. \tag{45}$$

For sufficiently large n, choose

$$m = \left\lfloor \left(\frac{n}{\log^3(en)} \right)^{1/(2\alpha+d)} \right\rfloor.$$

Then $m \geq 2$, m is within a constant factor of the quantity inside the floor, and $\log m \leq C \log(en)$. Substituting this choice into (45) gives

$$\mathrm{Err}_Q(\widehat{s}) \leq C \left(\frac{\log^3(en)}{n} \right)^{2\alpha/(2\alpha+d)} + C \frac{\log(1/\eta)}{n},$$

which proves the desired high-probability bound in the theorem. The architecture depends only on n and the fixed parameters, and the same choice works for every $K \geq 2$ and $N \geq 1$ with $n = NK$. □

C.3 Proof of the minimax lower bound

We prove Theorem 4.6 using a finite collection of joint distributions with common known uniform marginals. All constants in this subsection depend only on α, d_X, d_Y, M and on a fixed smooth function introduced below.

Proof of Theorem 4.6. Step 1. Choose a nonzero infinitely differentiable function ψ_0 supported inside $(1/4, 3/4)$ such that

$$\int_0^1 \psi_0(t)dt = 0, \qquad \|\psi_0\|_\infty \leq 1.$$

On $[0,1]^{d_X+d_Y}$, set $\psi(z) = \prod_{\ell=1}^{d_X+d_Y} \psi_0(z_\ell)$ and extend it by zero. For an integer $m \geq 1$, set $h = 1/m$, partition the unit cube into $m^{d_X+d_Y}$ cubes of side h, and let z_j denote the lower-left corner of the jth cube. Set $\psi_j(z) = \psi((z - z_j)/h)$. The supports are disjoint, and

$$\int \psi_j dQ = 0, \qquad \|\psi_j\|^2_{L^2(Q)} = h^{d_X+d_Y} \|\psi\|^2_{L^2},$$

where Q is the uniform law on the product cube. Integrating $\psi_j(x, y)$ over either modality also gives zero for every value of the other modality. For $\theta \in \{-1, 1\}^{m^{d_X+d_Y}}$, define the density

$$p_\theta(z) = 1 + \epsilon h^\alpha \sum_{j=1}^{m^{d_X+d_Y}} \theta_j \psi_j(z),$$

where $0 < \epsilon \leq 1/4$ is fixed below, independently of m and n. At most one summand is nonzero at any point, so $3/4 \leq p_\theta \leq 5/4$. The zero-integral properties imply that this is a probability density and that both its marginals are uniform. Let P_θ denote the corresponding distribution of (X, Y). Thus the product of the marginals is the same Q for every θ, and the PMI is exactly $s^\star_\theta = \log p_\theta$.

Step 2. Recall $r = \lceil \alpha \rceil - 1$ and $\beta = \alpha - r$. For $|t| \leq 1/4$, the chain rule and $1 + t\psi \geq 3/4$ imply

$$\|\partial^\nu \log(1 + t\psi)\|_\infty \leq C|t|, \qquad |\nu| \leq r + 1.$$

For order zero this follows from the mean value theorem. In every positive-order derivative at least one factor t occurs, while all derivatives of ψ and the inverse denominators are bounded. Substituting $t = \theta_j \epsilon h^\alpha$ and rescaling within cube j yields

$$\|\partial^\nu s^\star_\theta\|_\infty \leq C\epsilon h^{\alpha-|\nu|}, \qquad |\nu| \leq r + 1.$$

These are global bounds: the functions vanish near cube boundaries and join smoothly, and their supports do not overlap. For $|\nu| = r$, the bounds at orders r and $r + 1$ give

$$|\partial^\nu s^\star_\theta(z) - \partial^\nu s^\star_\theta(z')| \leq C\epsilon \min\{h^\beta, h^{\beta-1}\|z - z'\|_\infty\} \leq C\epsilon \|z - z'\|^\beta_\infty.$$

For the last step, use the second term if $\|z - z'\|_\infty \leq h$ and the first otherwise. Hence $\|s^\star_\theta\|_{C^\alpha} \leq C\epsilon$ for every m and θ. Fix $0 < \epsilon \leq \min\{1/4, M/C\}$. All constructed laws then belong to the class in the theorem.

Step 3. For an estimator $\widetilde{s} \in L^2(Q)$, estimate θ_j by the sign of $\int \widetilde{s}\psi_j dQ$, assigning $+1$ at zero, and denote the resulting estimate by $\widehat{\theta}_j$. On the support of ψ_j, set $t = \epsilon h^\alpha$. Since

$$\log(1 + \theta_j t\psi_j) = \theta_j t\psi_j \int_0^1 \frac{du}{1 + u\theta_j t\psi_j},$$

we have

$$\theta_j \int s^\star_\theta \psi_j dQ = t \int \psi_j^2 \int_0^1 \frac{du}{1 + u\theta_j t\psi_j} dQ \geq \frac{4}{5} t \|\psi_j\|^2_{L^2(Q)}.$$

If $\widehat{\theta}_j \neq \theta_j$, it follows that, for every constant c,

$$\left| \int (\widetilde{s} - s^\star_\theta - c)\psi_j dQ \right| \geq \frac{4}{5} t \|\psi_j\|^2_{L^2(Q)}.$$

The constant disappears because $\int \psi_j dQ = 0$. The normalized functions $\psi_j / \|\psi_j\|_{L^2(Q)}$ are orthonormal. Bessel's inequality, followed by the infimum over c, therefore gives

$$\mathrm{Err}_Q(\widetilde{s}) \geq \frac{16}{25} \epsilon^2 h^{2\alpha + d_X + d_Y} \|\psi\|^2_{L^2} \sum_{j=1}^{m^{d_X + d_Y}} \mathbf{1}\{\widehat{\theta}_j \neq \theta_j\}. \tag{46}$$

If the estimator is not in $L^2(Q)$ on a sample outcome, its error is infinite because the target is bounded. On such outcomes, define $\widehat{\theta}_j$ arbitrarily for each j. The same inequality then continues to hold, so the argument applies to every measurable estimator in the theorem.

Step 4. Let θ and θ' differ only in their jth coordinate. The inequality $\log u \leq u - 1$ and the fact that both densities integrate to one give

$$D_{\mathrm{KL}}(P_\theta \| P_{\theta'}) \leq \int \frac{(p_\theta - p_{\theta'})^2}{p_{\theta'}} dQ \leq \frac{16}{3} \epsilon^2 h^{2\alpha + d_X + d_Y} \|\psi\|^2_{L^2}.$$

Choose $m = \lceil n^{1/(2\alpha + d_X + d_Y)} \rceil$. Independence of the training pairs multiplies this divergence by n, so $D_{\mathrm{KL}}(P_\theta^{\otimes n} \| P_{\theta'}^{\otimes n}) \leq 1/3$, since $nh^{2\alpha + d_X + d_Y} \leq 1$ and $\epsilon \leq 1/4$, together with $\|\psi\|^2_{L^2} \leq 2^{-(d_X + d_Y)} \leq 1$ (as $|\psi_0| \leq 1$ is supported in $(1/4, 3/4)$). Pinsker's inequality then implies $\mathrm{TV}(P_\theta^{\otimes n}, P_{\theta'}^{\otimes n}) \leq 1/2$. For any estimator $\widehat{\theta}_j \in \{-1, 1\}$, the sum of its two error probabilities under these neighboring distributions is at least $1 - \mathrm{TV} \geq 1/2$. Indeed, if $\theta_j = +1$, $\theta'_j = -1$, and $A = \{\widehat{\theta}_j = +1\}$, then the sum of the two error probabilities is

$$P_\theta^{\otimes n}(A^c) + P_{\theta'}^{\otimes n}(A) = 1 - \{P_\theta^{\otimes n}(A) - P_{\theta'}^{\otimes n}(A)\} \geq 1 - \mathrm{TV}(P_\theta^{\otimes n}, P_{\theta'}^{\otimes n}).$$

Averaging over the two possible values of θ_j and over all remaining coordinates of θ gives

$$2^{-m^{d_X + d_Y}} \sum_\theta P_\theta^{\otimes n}\{\widehat{\theta}_j \neq \theta_j\} \geq \frac{1}{4} \qquad \text{for each } j.$$

Step 5. Take expectations in (46), average over θ, and use the testing bound above. For every estimator,

$$\sup_\theta \mathbb{E}_{P_\theta^{\otimes n}} \mathrm{Err}_Q(\widetilde{s}) \geq \frac{4}{25} \epsilon^2 \|\psi\|^2_{L^2} h^{2\alpha}.$$

For $n \geq 1$, the choice of m gives $m \leq 2n^{1/(2\alpha + d_X + d_Y)}$. The right-hand side of the above inequality is consequently at least $cn^{-2\alpha/(2\alpha + d_X + d_Y)}$ with $c > 0$ depending only on the fixed parameters. Taking the infimum over estimators proves the theorem. □

D Representation learning and zero-shot classification

In this section, we prove Propositions 5.1 and 5.2, which establish the approximate sufficiency and zero-shot classification guarantees in Section 5.

D.1 Proof of Proposition 5.1

Proof of Proposition 5.1. Let $s(x, y) = \langle f(x), g(y)\rangle/\tau$ be a CLIP-style score with encoders f, g satisfying $\|s\|_\infty, \|s^\star\|_\infty \le M_0$. Define

$$q_s(dy \mid x) = \frac{e^{s(x,y)}}{\int e^{s(x,v)} P_Y(dv)}\, P_Y(dy).$$

Since $s(x, y) = \langle f(x), g(y)\rangle/\tau$ depends on x only through $f(x)$, both the numerator and the normalizing factor in $q_s(\cdot \mid x)$ depend on x only through $f(x)$. Let $U = f(X)$. Thus $q_s(\cdot \mid X)$ can equivalently be written as $q_s(\cdot \mid U)$. We can then decompose the conditional density ratio as

$$\log \frac{dP_{Y|X}}{dq_s(\cdot \mid X)} = \log \frac{dP_{Y|X}}{dP_{Y|U}} + \log \frac{dP_{Y|U}}{dq_s(\cdot \mid U)}.$$

Taking expectations and conditioning the second term on U gives

$$\begin{aligned} \mathbb{E}_X D_{\mathrm{KL}}\big(P_{Y|X} \,\|\, q_s(\cdot \mid X)\big) &= \mathbb{E} \log \frac{dP_{Y|X}}{dP_{Y|U}} + \mathbb{E}_U D_{\mathrm{KL}}\big(P_{Y|U} \,\|\, q_s(\cdot \mid U)\big) \\ &= \mathrm{Suff}(f) + \mathbb{E}_U D_{\mathrm{KL}}\big(P_{Y|U} \,\|\, q_s(\cdot \mid U)\big) \\ &\ge \mathrm{Suff}(f). \end{aligned} \tag{47}$$

Indeed, because $U = f(X)$ is a function of X,

$$\mathbb{E} \log \frac{dP_{Y|X}}{dP_{Y|U}} = I(X; Y \mid U) = I(X; Y) - I(U; Y) = \mathrm{Suff}(f).$$

Set $\Delta = s - s^\star$. For P_X-almost every x, define

$$A_x(t) = \log \int e^{s^\star(x,y) + t\Delta(x,y)} P_Y(dy), \qquad 0 \le t \le 1.$$

Since

$$\frac{dP_{Y|X=x}}{dP_Y}(y) = e^{s^\star(x,y)}$$

for P_X-almost every x, we have $A_x(0) = 0$. Define the probability measure

$$q_{t,x}(dy) = e^{s^\star(x,y) + t\Delta(x,y) - A_x(t)} P_Y(dy).$$

Then $q_{0,x} = P_{Y|X=x}$ and $q_{1,x} = q_s(\cdot \mid x)$. Differentiation under the integral is justified by the boundedness condition and gives

$$A_x'(t) = \mathbb{E}_{q_{t,x}} \Delta(x, Y), \qquad A_x''(t) = \mathrm{Var}_{q_{t,x}}\{\Delta(x, Y)\}.$$

Moreover, $s^\star + t\Delta = (1 - t)s^\star + ts$, so its absolute value is bounded by M_0. Consequently,

$$\frac{dq_{t,x}}{dP_Y} = \frac{e^{s^\star(x,\cdot) + t\Delta(x,\cdot)}}{\int e^{s^\star(x,v) + t\Delta(x,v)} P_Y(dv)} \le e^{2M_0}.$$

Thus, for every constant c,

$$\begin{aligned} A_x''(t) &= \mathrm{Var}_{q_{t,x}}\{\Delta(x, Y)\} \\ &\le \int \{\Delta(x, y) - c\}^2 q_{t,x}(dy) \\ &\le e^{2M_0} \int \{\Delta(x, y) - c\}^2 P_Y(dy). \end{aligned}$$

Next, using $q_{1,x} = q_s(\cdot \mid x)$, we have

$$\log \frac{dP_{Y|X=x}}{dq_s(\cdot \mid x)}(y) = s^\star(x, y) - s(x, y) + A_x(1) = -\Delta(x, y) + A_x(1).$$

Therefore,

$$D_{\mathrm{KL}}\big(P_{Y|X=x} \,\|\, q_s(\cdot \mid x)\big) = A_x(1) - \mathbb{E}_{Y|X=x}\Delta(x,Y) = A_x(1) - A_x(0) - A_x'(0).$$

Taylor's formula now gives

$$D_{\mathrm{KL}}\big(P_{Y|X=x} \,\|\, q_s(\cdot \mid x)\big) = \int_0^1 (1-t)A_x''(t)dt \leq \frac{e^{2M_0}}{2}\int \{\Delta(x,y)-c\}^2 P_Y(dy).$$

Integrating over P_X, taking the infimum over the same constant $c \in \mathbb{R}$, and applying (47), we obtain

$$\mathrm{Suff}(f) \leq \frac{e^{2M_0}}{2}\inf_{c\in\mathbb{R}}\int \{s(x,y)-s^\star(x,y)-c\}^2 Q(dx,dy) = \frac{e^{2M_0}}{2}\,\mathrm{Err}_Q(s).$$

Applying this bound to $(s,f,g) = (\widehat{s},\widehat{f},\widehat{g})$ proves the claim for $\widehat{f}$. Interchanging X and Y, and f and g, gives the corresponding bound for $\widehat{g}$. □

D.2 Prompt coverage and proof of Proposition 5.2

In this subsection, we provide additional details for the zero-shot classification setting considered in Section 5.2. We first give a simple sufficient condition under which the logits based on the true PMI recover the downstream class probabilities. We then introduce the prompt coverage condition used to control the effect of PMI estimation error, and explain how it applies to common choices of prompt distributions, including the class-conditional distributions considered in Uesaka et al. (2025) and point-mass distributions corresponding to single prompt classification in CLIP (Radford et al., 2021). Finally, we prove Proposition 5.2.

When the PMI logits recover the downstream class probabilities. Recall from Section 5.2 that, for the true PMI $s^\star$, the logit for class c is

$$z_c(x;s^\star) = b_c + \int s^\star(x,y)\nu_c(dy).$$

On the other hand, the optimal logits for cross-entropy classification are given by $\log P(C=c \mid X=x)$, up to a term common to all classes. Therefore, the logits based on $s^\star$ recover the downstream class probabilities whenever there exists a function $a(x)$, independent of c, such that

$$b_c + \int s^\star(x,y)\nu_c(dy) = \log P(C=c \mid X=x) + a(x), \qquad c=1,\dots,m,$$

for P_X-almost every x. The common term $a(x)$ has no effect after applying the softmax.

To illustrate this, we give one simple sufficient condition corresponding to a special case of the conditions considered by Uesaka et al. (2025). Suppose that $A_1,\dots,A_m$ form a disjoint partition of the text space with $P_Y(A_c) > 0$, and let

$$\nu_c = P_Y(\cdot \mid A_c).$$

For example, A_c may collect text prompts describing class c, such as "a photo of a dog" or "a dog in the park" when c corresponds to the class "dog." Assume further that

$$P_{Y|X=x} = \sum_{c=1}^m P(C=c \mid X=x)\nu_c.$$

Thus, x affects the text distribution only through the downstream class probabilities, while the distribution of texts within each A_c remains fixed as ν_c. Since $\nu_c = P_Y(\cdot \mid A_c)$, for $y \in A_c$,

$$e^{s^\star(x,y)} = \frac{dP_{Y|X=x}}{dP_Y}(y) = \frac{P(C=c \mid X=x)}{P_Y(A_c)}.$$

Hence $s^\star(x,y)$ is constant over $y \in A_c$, and

$$\int s^\star(x,y)\nu_c(dy) = \log P(C=c \mid X=x) - \log P_Y(A_c).$$

Choosing $b_c = \log P_Y(A_c)$ therefore gives

$$z_c(x; s^\star) = \log P(C=c \mid X=x),$$

so the logits based on the true PMI recover the downstream class probabilities exactly.

The coverage condition. Our theoretical results control the score error in $L^2(Q)$, where $Q = P_X \otimes P_Y$, whereas the zero-shot logits involve averages with respect to the class-specific prompt distributions ν_c. To relate these two quantities, suppose that, for each class c, $\nu_c \ll P_Y$ and let

$$w_c = \frac{d\nu_c}{dP_Y}.$$

Assume

$$\widetilde{C} = \sum_{c=1}^{m} \int w_c(y)^2 P_Y(dy) = \sum_{c=1}^{m} \|w_c\|^2_{L^2(P_Y)} < \infty. \tag{48}$$

The constant $\widetilde{C}$ measures how well the prompt distributions are covered by the pretraining text marginal P_Y. In particular, $\|w_c\|^2_{L^2(P_Y)}$ is small when ν_c is not too concentrated relative to P_Y, and becomes large when ν_c places substantial mass on regions with small P_Y-probability.

Now observe that for any $u \in L^2(Q)$,

$$\int u(x,y)\nu_c(dy) = \int u(x,y) w_c(y) P_Y(dy).$$

By Cauchy-Schwarz,

$$\left(\int u(x,y)\nu_c(dy)\right)^2 \le \left(\int u(x,y)^2 P_Y(dy)\right)\left(\int w_c(y)^2 P_Y(dy)\right).$$

Summing over c and averaging over X therefore gives

$$\mathbb{E}_X \sum_{c=1}^{m} \left(\int u(X,y)\nu_c(dy)\right)^2 \le \widetilde{C}\|u\|^2_{L^2(Q)}. \tag{49}$$

The coverage condition (48) is satisfied by several common choices of prompt distributions. If A_c is a set of prompts with $P_Y(A_c) > 0$ and $\nu_c = P_Y(\cdot \mid A_c)$, as in Uesaka et al. (2025), then

$$\widetilde{C} = \sum_{c=1}^{m} \frac{1}{P_Y(A_c)}.$$

For single point-mass prompts $\nu_c = \delta_{y_c}$, corresponding to single prompt classification as in CLIP (Radford et al., 2021), the condition holds whenever $P_Y(\{y_c\}) > 0$ for every c, and

$$\widetilde{C} = \sum_{c=1}^{m} \frac{1}{P_Y(\{y_c\})}.$$

More generally, if each ν_c is an empirical distribution supported on finitely many points with positive P_Y-mass, then

$$\widetilde{C} = \sum_{c=1}^{m} \sum_{y} \frac{\nu_c(\{y\})^2}{P_Y(\{y\})}.$$

Proof of Proposition 5.2. Write $\Delta = \widehat{s} - s^\star$ and let $c_0 = \int \Delta dQ$. By the definition of Err_Q, $\|\Delta - c_0\|^2_{L^2(Q)} = \mathrm{Err}_Q(\widehat{s})$. Since each ν_c is a probability distribution, subtracting c_0 from the score

shifts every class logit by the same amount and therefore does not change the softmax probabilities or the cross-entropy. We may thus compare $z^\star(x) = z(x; s^\star)$ with $z^\star(x) + v(x)$, where

$$v_c(x) = \int \{\Delta(x, y) - c_0\}\nu_c(dy).$$

Applying (49) with $u = \Delta - c_0$ gives

$$\mathbb{E}_X \|v(X)\|_2^2 \le \widetilde{C}\,\mathrm{Err}_Q(\widehat{s}).$$

Now let $\pi_c = P(C = c \mid X = x)$ and $q_c = \mathrm{softmax}(z^\star(x))_c$. For $F(z) = \log \sum_{c=1}^m e^{z_c}$, writing $p = \mathrm{softmax}(z)$, direct differentiation gives $\nabla F(z) = p$ and $\nabla^2 F(z) = \mathrm{diag}(p) - pp^\top$. Hence, for every $u \in \mathbb{R}^m$,

$$u^\top \nabla^2 F(z) u = \mathrm{Var}_p(u) \le \frac{(\max_c u_c - \min_c u_c)^2}{4} \le \frac{1}{2}\|u\|_2^2.$$

Taylor's formula therefore gives $F(z^\star(x) + v) - F(z^\star(x)) - q^\top v \le \|v\|_2^2/4$. Using $D_{\mathrm{KL}}(\pi \| \mathrm{softmax}(z)) = \sum_{c=1}^m \pi_c \log \pi_c - \pi^\top z + F(z)$, with $0 \log 0 = 0$, we obtain

$$D_{\mathrm{KL}}\big(\pi \| \mathrm{softmax}(z^\star(x) + v)\big) \le D_{\mathrm{KL}}(\pi \| q) + (q - \pi)^\top v + \frac{1}{4}\|v\|_2^2.$$

Pinsker's inequality gives $\|q - \pi\|_2 \le \|q - \pi\|_1 \le \sqrt{2 D_{\mathrm{KL}}(\pi \| q)}$. Therefore,

$$(q - \pi)^\top v \le \sqrt{2 D_{\mathrm{KL}}(\pi \| q)}\|v\|_2 \le D_{\mathrm{KL}}(\pi \| q) + \frac{1}{2}\|v\|_2^2.$$

Combining this with the bound above yields

$$D_{\mathrm{KL}}\big(\pi \| \mathrm{softmax}(z^\star(x) + v)\big) \le 2 D_{\mathrm{KL}}(\pi \| q) + \frac{3}{4}\|v\|_2^2.$$

Averaging over X and using $\mathbb{E}_X D_{\mathrm{KL}}(\pi \| q) = R(z(\cdot; s^\star)) - R^\star$ and $\mathbb{E}_X \|v(X)\|_2^2 \le \widetilde{C}\,\mathrm{Err}_Q(\widehat{s})$ gives

$$R(z(\cdot; \widehat{s})) - R^\star \le 2\{R(z(\cdot; s^\star)) - R^\star\} + \frac{3}{4}\widetilde{C}\,\mathrm{Err}_Q(\widehat{s}),$$

which proves the proposition. □

E Additional experimental details and results

This section provides additional details for the synthetic experiments in Section 6 and presents additional experiments on real image-text data.

E.1 Additional details for synthetic experiments

We provide additional implementation and evaluation details for the synthetic experiments in Section 6. For the experiment with the joint-input ReLU score, we set $\lambda = 1.782$ so that $\mathrm{Var}_Q(s^\star) = 7.94$, and use $M_0 = 2\lambda d_X$ in the bounded score $s_\theta = M_0 \tanh(h_\theta / M_0)$. For the experiment with the normalized CLIP-style score, we set $\omega = 8.897$, which is chosen so that $\mathbb{E}[U^\top V] = 0.6$. The fixed nonlinear transformations h_X and h_Y used to generate the observed variables are of the form

$$h(z) = A_2\,\mathrm{LeakyReLU}(A_1 z + b_1) + b_2,$$

where the negative slope of the LeakyReLU activation is 0.2 and A_1 and A_2 are invertible matrices. We independently generate the transformations for the two modalities and keep them fixed throughout all experiments.

For both experiments, we use five independent trials. In each trial, we generate 2^{17} paired observations and, for each smaller value of n, use the first n observations, so that the datasets are nested as n increases. For each (n, K), consecutive observations are partitioned into fixed batches of size K. We train the model from three random initializations and retain the one with the smallest final empirical InfoNCE loss. Adam is used with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. To evaluate the learned score, we estimate $\mathrm{Err}_Q(\widehat{s})$ using 2^{18} independent samples from Q and report the relative error $\mathrm{Err}_Q(\widehat{s})/\mathrm{Var}_Q(s^\star)$. We also estimate the population excess InfoNCE risk $\mathcal{E}_K(\widehat{s}) = L_K(\widehat{s}) - L_K(s^\star)$ using a separate set of 2^{18} fresh paired observations from P, partitioned into batches of size K. The curves in Figure 1 report the mean and standard deviation over the five independent trials.

E.2 Experiments on real image-text data

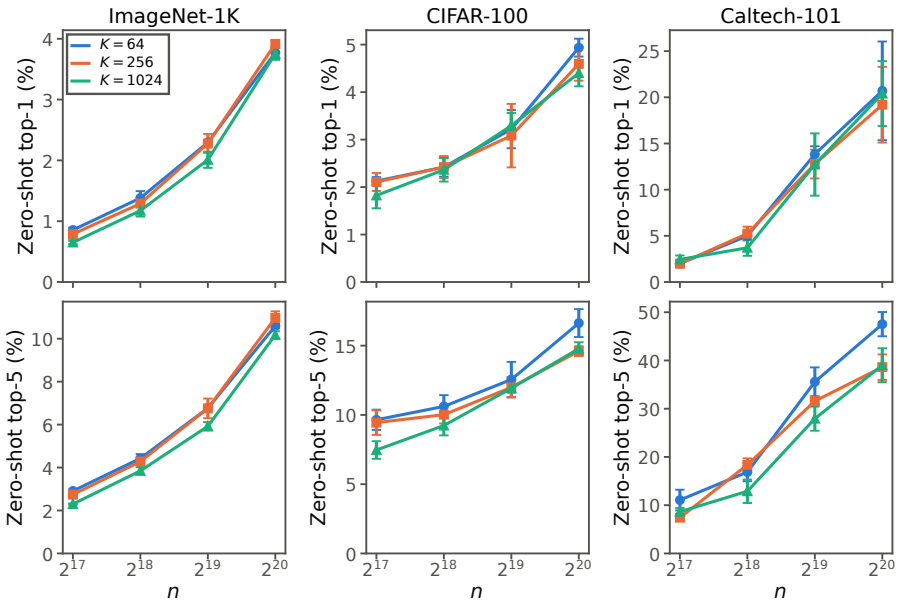

Figure 2: Zero-shot top-1 (top) and top-5 (bottom) classification accuracy on ImageNet-1K, CIFAR-100, and Caltech-101 versus the total number of paired pretraining samples n for different batch sizes K. The pretrained CLIP-style encoders are frozen and evaluated on each downstream classification task without further training. Curves show the mean with standard deviation over three training runs with different random seeds.

We use the English portion of the Wikipedia-based Image Text (WIT) dataset (Srinivasan et al., 2021) for pretraining, using the English reference description associated with each image as its caption. After removing observations with empty captions, unreadable images, and duplicate images, the resulting dataset contains approximately 2.36M image-text pairs. We use approximately 50,000 image-text pairs each for validation and testing, assigning all pairs associated with the same Wikipedia page to the same split, and use the remaining pairs for training. We randomly order the training pairs once and, for each $n \in \{2^{17}, 2^{18}, 2^{19}, 2^{20}\}$, use the first n pairs in this fixed order. Thus, the training sets are nested as n increases. We consider batch sizes $K \in \{64, 256, 1024\}$. For each (n, K), the training observations are partitioned into fixed disjoint batches of size K, while the order of the batches is randomly shuffled at each epoch.

For pretraining, we use a normalized CLIP-style score with fixed temperature $\tau = 0.07$ and train both encoders from scratch for 30 epochs. The image encoder is a ResNet-18 followed by a linear projection to dimension 256, while the text encoder is a 6-layer Transformer of width 256 with 4 attention heads, followed by a linear projection to dimension 256. Images are resized and center-cropped to 224×224 pixels without data augmentation. We optimize the InfoNCE loss using AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.98$, weight decay 0.1, and a cosine learning-rate schedule with 5% linear warm-up. The learning rate is set to $10^{-3}\sqrt{K/256}$ for all batch sizes K.

For downstream evaluation, we consider zero-shot classification on ImageNet-1K (Deng et al., 2009; Russakovsky et al., 2015), CIFAR-100 (Krizhevsky, 2009), and Caltech-101 (Fei-Fei et al., 2004). The encoders are frozen, and for each class we construct a text representation by averaging the normalized embeddings of standard prompt templates for that class. We use 80, 18, and 34 prompt templates per class for ImageNet-1K, CIFAR-100, and Caltech-101, respectively, and report top-1 and top-5 classification accuracy.

Figure 2 shows the results from the final training checkpoint, averaged over three training runs with different random seeds. Across all three downstream datasets, classification accuracy generally improves as the total number of pretraining samples n increases. At the same time, the curves corresponding to different batch sizes remain relatively close, particularly on ImageNet-1K and

CIFAR-100, while Caltech-101 exhibits somewhat larger differences across batch sizes. Overall, these results indicate that downstream performance improves primarily with the total number of pretraining samples and is comparatively less sensitive to the batch size.