POPULATION SIMILARITY ENABLES STATISTICALLY GROUNDED COMPARISONS OF REPRESENTATION ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Representation similarity metrics are widely used to evaluate neural network alignment. Owing to their quadratic time complexity, these metrics are frequently computed on small samples ($N \approx 1{,}000$), yielding empirical scores that can misrepresent the underlying data-generating process. To capture true alignment, we propose population similarity, i.e., the infinite-sample limit of an empirical score, as the primary target of analysis. We establish the existence of this limit and the consistency of the Triplet and Quadruplet Similarity Indices (TSI and QSI) and Centered Kernel Alignment with an RBF kernel (CKA-RBF), and show that Mutual Nearest Neighbors (MutualNN) admits a population similarity under a mild monotonicity condition. To quantify finite-sample deviations without distribution assumptions, we introduce distribution-free tail bounds, proving non-trivial bounds for TSI, QSI, and CKA-RBF, but not for MutualNN, implying its arbitrarily slow convergence and the failure of empirical scores to reflect its population score. We demonstrate experimentally that the ordering of representation pairs by their finite-sample scores is often inconsistent with their ordering by population similarity, with disagreement rates heavily dependent on the chosen metric and the representation pairs compared. Building on our bounds, we design a one-sided hypothesis test to determine if one population similarity exceeds another by a specified margin, and show its p-value can be upper-bounded using only observed scores and sample sizes. These bounds are informative for TSI and QSI at practical sample sizes ($1{,}000$ to $10{,}000$). Finally, we showcase the practical utility of our approach with a multimodal case study revealing that CLIP's image and text representations are less aligned at the population level than those of SigLIP2.

1 INTRODUCTION

There has been growing interest in analyzing the similarity between latent representations of deep neural networks (Sucholutsky et al., 2025). To achieve this, practitioners often use similarity metrics that take two sets of paired representations and output a scalar quantifying their similarity (Klabunde et al., 2025). This methodology has been used to study questions such as whether representations converge to a shared "platonic" representation (Huh et al., 2024; Gröger et al., 2026; Koepke et al., 2026), and whether generative models can leverage knowledge distillation to improve generation quality (Yu et al., 2025; Leng et al., 2025).

In this paper, we argue that *population similarity*, i.e., the infinite-sample limit of an empirical similarity score, is the quantity that captures true alignment under that similarity metric. An empirical score $S(X, Y)$ is a random variable that depends on the sampled representation set of size N, so it might misrepresent alignment due to sampling variability. In contrast, its population counterpart follows deterministically from the data-generating distribution and the chosen metric, invariant to N and to any specific sample. Still, despite their widespread use, analyses based on similarity metrics rarely account for the effect of N, which can lead to incorrect conclusions. This issue is especially acute because many similarity metrics have quadratic time complexity, forcing practitioners to evaluate scores on small subsets (Nguyen et al., 2021) that are often as small as a few thousand samples (Huh et al., 2024). At this scale, practitioners take an empirical ranking such as $S(X, Y) = 0.68 < S(X', Y') = 0.71$ as evidence that (X', Y') is more aligned than (X, Y) under

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.

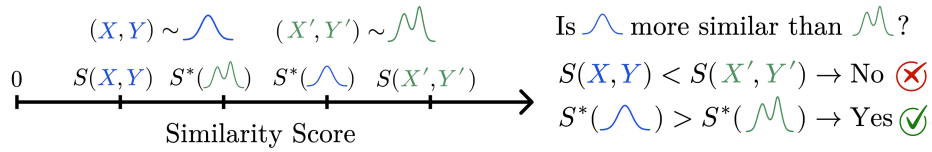

Figure 1: Empirical similarity scores can reverse the ranking of the corresponding population similarities, leading to incorrect conclusions about which pair of representations is more similar.

S, yet the *population rank*, i.e., the ordering induced by the corresponding population similarities, can be the reverse (see Figure 1). More generally, it is unclear what margin between the empirical scores would be sufficient to determine the population rank and how it depends on the choice of similarity score. As such, instead of directly comparing empirical similarity scores for a finite N, we propose using these as estimates of their corresponding population similarity scores. In doing so, we strengthen the reliability of such analyses, moving from sample-dependent random variables to well-defined quantities that capture the similarity of the underlying data-generating process.

However, in practice we cannot collect infinite samples, making direct computation of population similarity scores impossible. Therefore, to make reliable statements about the population ranks, we need a notion of how close empirical similarity scores are to their population counterparts. Here, we show that empirical scores for metrics such as CKA-RBF (Kornblith et al., 2019), TSI, and QSI (Soares et al., 2026) predictably approximate their population limits under non-trivial convergence bounds that are independent of the underlying data distribution. This distribution independence is essential because the true distribution is unknown and infinitely many distributions could generate any observed sample, so only distribution-free bounds can assess how reliably an empirical score approximates the population similarity.

Building on these bounds, we design a hypothesis testing framework to determine whether an observed empirical gap reflects a true population-level difference. As illustrated in Figure 2, this yields a computationally cheap analytical upper bound on the test's p-value, enabling practitioners to rigorously evaluate population-level hypotheses directly from finite samples without distributional assumptions. In short, our main contributions are:

1. **Population Similarity and Consistency:** We motivate the study of population similarity and prove that any bounded metric converging in probability, including TSI, QSI and CKA-RBF, admits a population similarity and is consistent.
2. **Distribution-Free Tail Bounds:** We define distribution-free tail bounds on the convergence rate of empirical scores towards their population similarities. We derive non-trivial bounds for TSI, QSI, and CKA-RBF, and prove that no such bounds exist for MutualNN.
3. **One-Sided Comparison Test and P-value Upper Bounds:** We formalize a one-sided hypothesis test to evaluate whether one population similarity exceeds another by a specified margin $t \geq 0$. For any metric admitting distribution-free tail bounds, we prove that the test's p-value can be analytically upper-bounded using solely the observed empirical scores and sample sizes.
4. **Practical Insights:** We translate these guarantees into actionable takeaways for practitioners: (i) we show that metrics lacking distribution-free bounds, including MutualNN, can converge arbitrarily slowly to their population counterpart; and (ii) we show that our p-value bounds are practically informative for TSI and QSI at accessible sample sizes (1,000 to 10,000), providing a useful tool for robust population-level comparisons.
5. **Empirical Findings:** We evaluate our framework across multimodal representations: (i) empirical similarity scores misrepresent population limits at standard batch sizes; (ii) finite-sample rank disagreement rates vary sharply across metrics and model pairs, making raw empirical rankings unreliable; and (iii) our hypothesis test confirms with statistical confidence that SigLIP2's image and text representations are strictly more aligned at the population level than CLIP's.

2 Related Work

Similarity Metrics. A wide range of metrics have been proposed to quantify representation alignment (Sucholutsky et al., 2025; Klabunde et al., 2025). Among these, Centered Kernel Alignment

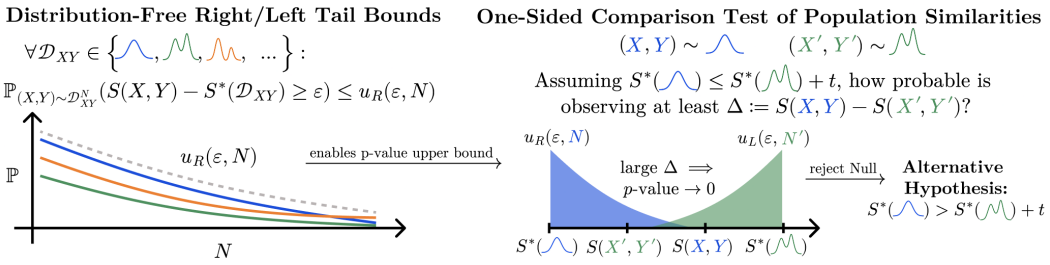

Figure 2: Similarity metrics admitting distribution-free tail bounds (left) enable statistically grounded comparisons of population similarities via a one-sided test (right), using only empirical scores and sample sizes.

(CKA) (Kornblith et al., 2019) remains widely adopted, evaluating alignment through the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2005). More recently, local and ordinal geometries have motivated alternative formulations: Mutual Nearest Neighbors (MutualNN) (Huh et al., 2024) evaluates shared k-NN graph overlap, while the Triplet (TSI) and Quadruplet (QSI) Similarity Indices (Soares et al., 2026) measure the ordinal relationships. While these metrics are widely used to analyze representational alignment (Huh et al., 2024; Yu et al., 2025; Koepke et al., 2026), their theoretical properties, such as asymptotic population limits and finite-sample convergence behavior, have received relatively less systematic study.

Statistical Reliability and Calibration of Similarity Scores. Growing awareness of the sensitivity of representation metrics has prompted several lines of work on their statistical reliability (Ding et al., 2021; Davari et al., 2023). A prominent issue is that raw similarity scores often lack a meaningful baseline, as unrelated representations can yield deceptively high empirical similarity (Cui et al., 2022). Gröger et al. (2026) address this by calibrating scores against a permutation-derived null distribution, in a similar spirit to the nonparametric permutation and bootstrap procedures long used in representational similarity analysis (Nili et al., 2014). Ding et al. (2021) instead use bootstrap resampling to construct confidence intervals on rank-correlation differences between metrics, and Ciernik et al. (2025) empirically characterize how many samples CKA variants require to stabilize on large-scale vision benchmarks. While the population limit is known to exist for some metrics, such as centered kernel alignment for bounded kernels (Cortes et al., 2012) or ordinal indices (Soares et al., 2026), practical reliability assessments still largely rely on empirical resampling, which is inherently conditioned on the observed finite sample. Because the true data-generating distribution is unknown, resampling cannot substitute for analytical finite-sample guarantees, nor does it typically offer a unified framework across different metrics.

3 Notation and Background

Let $\mathcal{X}$ and $\mathcal{Y}$ be representation spaces equipped with distance functions d_X and d_Y, and let $\mathcal{D}_{XY}$ be a joint distribution over pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$. We write $(X, Y) \sim \mathcal{D}_{XY}^N$ for an i.i.d. sample of N pairs, with $X = (x_1, \ldots, x_N)$ and $Y = (y_1, \ldots, y_N)$. Furthermore, we denote *similarity metrics* as functions S_{d_X,d_Y} mapping a paired sample (X, Y) to a score in $[0, 1]$. Finally, we say that $S_{d_X,d_Y}(X, Y)$ *converges in probability* to a constant $c \in \mathbb{R}$ if, for every $\varepsilon > 0$, $\lim_{N\to\infty} \mathbb{P}_{(X,Y)\sim\mathcal{D}_{XY}^N}\big(|S_{d_X,d_Y}(X, Y) - c| \geq \varepsilon\big) = 0$, and write $S_{d_X,d_Y}(X, Y) \xrightarrow[N\to\infty]{\mathbb{P}} c$. In this work, we study four similarity metrics (Definitions 1 to 3).

Definition 1 (Triplet and Quadruplet Similarity Indices (TSI, QSI) (Soares et al., 2026))**.** TSI and QSI are built from the ordinal comparison $O_d(a, b, c, d) = \operatorname{sign}\big(d(a, b) - d(c, d)\big)$, with $O_d^\Delta(a, b, c) \equiv O_d(a, b, a, c)$. Let $\mathcal{T}$ (resp. $\mathcal{Q}$) be the set of all ordered triplets (resp. quadruplets) of distinct indices in $\{1, \ldots, N\}$. Then

$$\mathrm{TSI}_{d_X,d_Y}(X, Y) = \mathbb{E}_{(i,j,k)\in\mathcal{T}}\Big[\mathbb{I}\big[O_{d_X}^\Delta(x_i, x_j, x_k) = O_{d_Y}^\Delta(y_i, y_j, y_k)\big]\Big],$$

$$\mathrm{QSI}_{d_X,d_Y}(X, Y) = \mathbb{E}_{(i,j,k,l)\in\mathcal{Q}}\Big[\mathbb{I}\big[O_{d_X}(x_i, x_j, x_k, x_l) = O_{d_Y}(y_i, y_j, y_k, y_l)\big]\Big].$$

Definition 2 (Centered Kernel Alignment with RBF Kernel (CKA-RBF)(Kornblith et al., 2019))**.** Let $\sigma_X, \sigma_Y > 0$ be fixed bandwidths. Let $k_X(x, x') = \exp(-d_X(x, x')^2/2\sigma_X^2)$ be the RBF kernel on $\mathcal{X}$, with Gram matrix $(K_X)_{ij} = k_X(x_i, x_j)$, and k_Y, K_Y defined analogously. Define

$\mathrm{HSIC}_{\sigma_X,\sigma_Y,d_X,d_Y}(X,Y) = \frac{1}{(N-1)^2}\operatorname{tr}(K_X H K_Y H)$ with centering matrix $H = I_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$, where I_N is the $N \times N$ identity and $\mathbf{1}\mathbf{1}^\top$ is the $N \times N$ all-ones matrix.

$$\text{CKA-RBF}_{\sigma_X,\sigma_Y,d_X,d_Y}(X,Y) = \frac{\mathrm{HSIC}_{\sigma_X,\sigma_Y,d_X,d_Y}(X,Y)}{\sqrt{\mathrm{HSIC}_{\sigma_X,\sigma_X,d_X,d_X}(X,X)\,\mathrm{HSIC}_{\sigma_Y,\sigma_Y,d_Y,d_Y}(Y,Y)}}.$$

Definition 3 (Mutual Nearest Neighbors (MutualNN)(Huh et al., 2024))**.** For $k \in \mathbb{N}$ and a sequence $Z = (z_1, \ldots, z_N)$, let $\mathrm{KNN}_{k,d}(Z,i)$ denote the indices of the k nearest neighbors of z_i among $Z \setminus \{z_i\}$, measured by distance d (breaking ties arbitrarily).

$$\mathrm{MutualNN}_{k,d_X,d_Y}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}\frac{|\mathrm{KNN}_{k,d_X}(X,i) \cap \mathrm{KNN}_{k,d_Y}(Y,i)|}{k}.$$

The metrics above are well-defined under Assumption 1. For the remaining regularity conditions assumed throughout, see Appendix A.

Assumption 1 (Well-defined similarity metrics)**.** Throughout, all metrics are assumed well-defined with admissible sample sizes: $N \geq 3$ for TSI, $N \geq 4$ for QSI, $N \geq 2$ for CKA-RBF, and $N \geq k+1$ for $\mathrm{MutualNN}_k$, as well as minor regularity conditions (Assumption 2). Additionally, for $\text{CKA-RBF}_{\sigma_X,\sigma_Y}$ we assume that there exists a uniform constant $m > 0$ such that for all admissible distributions, distances, and sample sizes:

$$\mathbb{P}_{(X,Y)\sim\mathcal{D}_{XY}^N}\left(\mathrm{HSIC}_{\sigma_X,\sigma_X}(X,X) \geq m \text{ and } \mathrm{HSIC}_{\sigma_Y,\sigma_Y}(Y,Y) \geq m\right) = 1.$$

This assumption ensures that the denominator of CKA-RBF is well-behaved by ruling out degenerate representations where it vanishes or becomes arbitrarily small. These conditions often hold in practice. Because the Gram matrices are centered, vanishing self-HSIC requires the kernel to assign near-constant values across representation pairs. This can only happen if the data distribution lacks diversity, i.e. frequently yielding identical representations, or if the kernel bandwidth fails to resolve differences between distinct points. Neither scenario is expected to occur in non-degenerate representation analyses.

4 Comparing Representation Alignment at the Population Level

As motivated in Section 1, we shift representation alignment from empirical scores to population similarities. This section asks whether, for a given metric, population similarity exists and whether empirical scores converge to it. When they do, we derive *distribution-free* bounds on the sample size N needed to approximate this quantity arbitrarily closely, holding across any data distribution. Finally, we use these bounds to quantify the confidence of a comparison between two population similarities from the observed empirical scores and sample sizes.

4.1 Formal Definitions & Existence Guarantees

The key object of this work is the expected similarity score in the large-sample limit.[1]

Definition 4 (Population Similarity)**.** For a similarity metric S, we define its *Population Similarity* (or true similarity) as:

$$S^*_{d_X,d_Y}(\mathcal{D}_{XY}) = \lim_{N\to\infty}\mathbb{E}_{(X,Y)\sim\mathcal{D}_{XY}^N}\left[S_{d_X,d_Y}(X,Y)\right].$$

Existence of this limit does not imply that empirical scores converge to it, as the expectation may converge while scores remain far from S^*. For instance, a bounded metric that depends only on the first pair (x_1, y_1) and is non-constant in that pair has a well-defined population similarity, yet $S(X,Y)$ might not converge in probability to S^*. We capture this distinction with consistency.

Definition 5 (Consistent Similarity Metric)**.** A similarity metric S is *consistent* if for every joint probability distribution $\mathcal{D}_{XY}$ and distance functions d_X, d_Y for which the population similarity $S^*_{d_X,d_Y}(\mathcal{D}_{XY})$ exists, and every $\varepsilon > 0$,

$$\lim_{N\to\infty}\mathbb{P}_{(X,Y)\sim\mathcal{D}_{XY}^N}\left(|S_{d_X,d_Y}(X,Y) - S^*_{d_X,d_Y}(\mathcal{D}_{XY})| \geq \varepsilon\right) = 0.$$

[1] This quantity was introduced as population similarity by Soares et al. (2026), but not further explored.

Does Population Similarity Always Exist? For population similarity to be a usable framework, it must be well-defined. By Definition 4, this requires that $\mathbb{E}[S_{d_X,d_Y}(X,Y)]$ exists for each N, and that these expectations have a limit as $N \to \infty$.

Theorem 1 (Existence of Population Similarity)**.** *Let S be a bounded similarity metric, i.e., $\mathrm{Im}(S) \subseteq [0,1]$, and suppose that for every joint distribution $\mathcal{D}_{XY}$ and distance functions d_X, d_Y there exists $c \in [0,1]$ such that $S_{d_X,d_Y}(X,Y) \xrightarrow[N\to\infty]{\mathbb{P}} c$ for $(X,Y) \sim \mathcal{D}_{XY}^N$. Then, for every $\mathcal{D}_{XY}, d_X, d_Y$, $S^*_{d_X,d_Y}(\mathcal{D}_{XY})$ exists in $[0,1]$ and equals the in-probability limit of $S_{d_X,d_Y}(X,Y)$. In particular, S is consistent.*

This result (Theorem 1) establishes existence and consistency under mild conditions that any meaningful similarity metric, existing or yet to be developed, is expected to satisfy from first principles. First, boundedness is fundamental because any sensible measure of alignment requires well-defined extrema, such that identity achieves maximal similarity while independent representations attain the theoretical minimum, which holds for most metrics in the literature (Klabunde et al., 2025; Gröger et al., 2026). Restricting that range to $[0,1]$ is typically without loss of generality, as scores in $[a,b]$ map to $[0,1]$ via a rank-preserving affine transformation. Second, convergence in probability ensures that similarity reflects the data generating process rather than sample artifacts; as the empirical distribution converges to the underlying population, any well-behaved metric should naturally stabilize around a unique, deterministic value. Consequently, our existence guarantees inherently apply to the vast majority of practical similarity metrics, including TSI, QSI and CKA-RBF, which we instantiate next.

Corollary 1 (Existence of Population Similarity and Consistency of Similarity Metrics)**.** *Under Assumption 1, for every joint distribution $\mathcal{D}_{XY}, d_X, d_Y$,* TSI*,* QSI*, and* $\mathrm{CKA\text{-}RBF}_{\sigma_X,\sigma_Y}$ *have population similarities and are consistent.* $\mathrm{MutualNN}_k$ *has a population similarity if, for all sufficiently large N, $\mathbb{E}_{(X,Y)\sim\mathcal{D}_{XY}^N}[\mathrm{MutualNN}_k(X,Y)] \geq \mathbb{E}_{(X,Y)\sim\mathcal{D}_{XY}^{N+1}}[\mathrm{MutualNN}_k(X,Y)]$.*

See Appendix E.2 for the proofs. For TSI, QSI, and CKA-RBF, existence and consistency follow together from convergence in probability (Theorem 1). For MutualNN we use a different monotonicity argument, which we expect to hold in practice, since adding points makes it harder for any point's k nearest neighbors to remain unchanged. The condition can nevertheless fail.

4.2 Analyzing the Convergence of Empirical Estimates

While establishing the existence of a population similarity and the consistency of its estimator is necessary, it is insufficient to draw reliable conclusions from finite samples. Practical representation analysis demands explicit convergence rates that quantify how rapidly an empirical score approximates S^*. Characterizing this rate is difficult because it typically depends on the underlying data distribution and distance functions, which are usually unknown. To address this limitation, we derive *distribution-free* tail bounds on finite-sample deviations that hold universally across all data distributions and admissible distances, providing reliable guarantees without unverifiable assumptions.

Definition 6 (Distribution-Free Tail Bounds)**.** Let S be a similarity metric. Rate functions $u_R, u_L : \mathbb{R}_+ \times \mathbb{N} \to [0,1]$ are *distribution-free right-tail* and *left-tail bounds* for S if, for every $\mathcal{D}_{XY}, d_X, d_Y$, $\varepsilon > 0$, and $N \in \mathbb{N}$,

$$\mathbb{P}_{(X,Y)\sim\mathcal{D}_{XY}^N}\big(S_{d_X,d_Y}(X,Y) - S^*_{d_X,d_Y}(\mathcal{D}_{XY}) \geq \varepsilon\big) \leq u_R(\varepsilon, N),$$
$$\mathbb{P}_{(X,Y)\sim\mathcal{D}_{XY}^N}\big(S^*_{d_X,d_Y}(\mathcal{D}_{XY}) - S_{d_X,d_Y}(X,Y) \geq \varepsilon\big) \leq u_L(\varepsilon, N).$$

A bound $u \in \{u_R, u_L\}$ is *non-trivial* if it is non-increasing in each argument and, for every $\varepsilon > 0$, satisfies $\lim_{N\to\infty} u(\varepsilon, N) = 0$.

Equipped with Definition 6, we can now characterize the finite-sample concentration behavior of our candidate metrics. As established in Corollary 2, TSI, QSI, and CKA-RBF concentrate exponentially around their population limits with rates that are invariant to the underlying distribution and distance functions. In contrast, MutualNN admits no such guarantees.

Corollary 2 (Existence of Distribution-Free Tail Bounds). *Under Assumption 1,* TSI, QSI, *and* CKA-RBF$_{\sigma_X,\sigma_Y}$ *admit the non-trivial distribution-free right- and left-tail bounds*

$$\begin{aligned} \text{TSI:} \quad & u_R(\varepsilon, N) = u_L(\varepsilon, N) = \exp\big(-2\lfloor N/3 \rfloor \varepsilon^2\big), \\ \text{QSI:} \quad & u_R(\varepsilon, N) = u_L(\varepsilon, N) = \exp\big(-2\lfloor N/4 \rfloor \varepsilon^2\big), \\ \text{CKA-RBF:} \quad & u_R(\varepsilon, N) = u_L(\varepsilon, N) = \min\big(1,\, 6\exp\big(-m^2 N \varepsilon^2/(64 \cdot 18^2)\big)\big), \end{aligned}$$

where $m > 0$ is the constant from Assumption 1. MutualNN$_k$ *admits no such bounds.*

See Appendix E.3 for the proofs. The theoretical implications of this result are significant. The sub-Gaussian tail bounds for TSI, QSI, and CKA-RBF ensure that observed deviations shrink predictably with N, providing an analytical foundation for comparing empirical scores to infer population rankings. Crucially, TSI and QSI require no distributional or geometric assumptions, making them strictly distribution-free. A caveat for CKA-RBF is that the bound applies only in the setting of Assumption 1. We describe it as distribution-free to signal the contrast with MutualNN, but this is a slight inaccuracy as pathological distributions or distance geometries can force the self-alignment constant m arbitrarily close to zero, rendering the bound distribution-dependent in that limit. Such cases are arguably ill-defined, since as $m \to 0$ the normalizing denominator vanishes and the kernel is nearly constant across pairs. In contrast, for MutualNN, the implications are far less reassuring.

Empirical MutualNN Scores can Converge Arbitrarily Slowly to its Population Similarity. To illustrate why MutualNN$_k$ admits no distribution-free tail bounds, consider the synthetic construction detailed in Appendix E.3.4. We construct pairs of representation spaces where the distances are designed so that the vast majority of points appear perfectly aligned. However, with a rare probability $\beta \in (0, 1)$, sampled points trigger a structural misalignment. Because the nearest-neighbor ranks of the entire set can be disrupted by just a few such points, the population similarity across the full distribution is strictly 0. Yet, on finite samples where these rare disruptive points have not yet been drawn, the empirical score remains deceptively close to 1. Evaluating MutualNN$_5$ averaged over 20 independent runs across varying sample sizes N (Figure 3), we observe that tuning β arbitrarily delays convergence. This exemplifies the fundamental issue highlighted by Corollary 2: without distribution-free tail bounds, convergence can be arbitrarily slow. Consequently, MutualNN is fundamentally unreliable for finite-sample analyses, as one cannot know for any given sample size whether an observed score reflects true underlying alignment or simply an unrevealed slow-convergence regime.

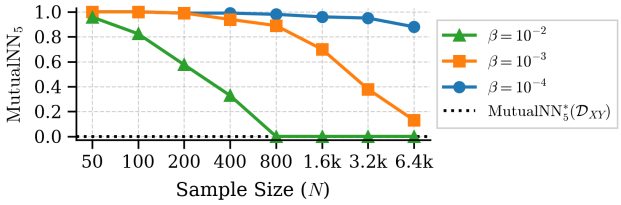

Figure 3: Empirical MutualNN scores can indicate strong alignment for arbitrarily large samples even when the representations are misaligned at the population level.

4.3 Statistical Significance of Comparison of Empirical Scores

Returning to our stated goal of comparing population similarities, suppose we observe empirical scores $S(X, Y) = 0.82$ and $S(X', Y') = 0.79$ on $N = N' = 1{,}000$ samples yielding an apparent gap of $\Delta = 0.03$. Due to finite-sample variance, this observed difference alone does not guarantee a consistent population ranking. To determine whether $S^*(\mathcal{D}_{XY})$ genuinely exceeds $S^*(\mathcal{D}_{X'Y'})$ by at least a specified margin $t \geq 0$, we treat the empirical difference as our test statistic and bound the tail probability of observing a gap of at least Δ under the unfavorable population ordering. We formalize this decision problem via the following one-sided hypothesis test.

Definition 7 (One-Sided Comparison Test of Population Similarities). Let S be a similarity metric, $t \geq 0$, and let $(X, Y) \sim \mathcal{D}_{XY}^N$ and $(X', Y') \sim \mathcal{D}_{X'Y'}^{N'}$ be samples with distances d_X, d_Y and $d_{X'}, d_{Y'}$. The associated test is

$$\begin{aligned} \textbf{Null Hypothesis } (H_0): & \; S^*_{d_X,d_Y}(\mathcal{D}_{XY}) \leq S^*_{d_{X'},d_{Y'}}(\mathcal{D}_{X'Y'}) + t, \\ \textbf{Alternative Hypothesis } (H_1): & \; S^*_{d_X,d_Y}(\mathcal{D}_{XY}) > S^*_{d_{X'},d_{Y'}}(\mathcal{D}_{X'Y'}) + t, \end{aligned}$$

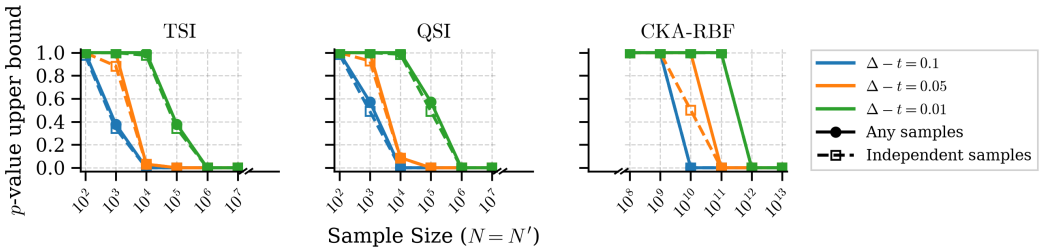

Figure 4: The sub-Gaussian distribution-free bounds of TSI, QSI and CKA (Corollary 2) yield an upper bound on the p-value by evaluating the infima in Theorem 2. These bounds are useful in practice for TSI and QSI.

with test statistic $T_S = S_{d_X,d_Y}(X,Y) - S_{d_{X'},d_{Y'}}(X',Y')$. For an observed value $T_S = \Delta$, significance is given by the tail probability

$$\mathbb{P}_{H_0}(T_S \geq \Delta) \coloneqq \mathbb{P}_{\substack{(X,Y)\sim\mathcal{D}_{XY}^N \\ (X',Y')\sim\mathcal{D}_{X'Y'}^{N'}}} \left(S_{d_X,d_Y}(X,Y) - S_{d_{X'},d_{Y'}}(X',Y') \geq \Delta\right),$$

where $\mathcal{D}_{XY}$ and $\mathcal{D}_{X'Y'}$ satisfy H_0.

Ideally, one would directly compute the exact tail probability (p-value) defined above. However, this is generally intractable because the data-generating distributions $\mathcal{D}_{XY}$ and $\mathcal{D}_{X'Y'}$ are unknown and the impact of the null configuration H_0 to the tail probability is unclear. Still, for any similarity metric admitting the distribution-free tail bounds established earlier, we can rigorously upper-bound this p-value using only the observed empirical gap, margin t, and sample sizes.

Theorem 2 (P-value Bounds from Distribution-Free Tail Bounds)**.** *Let S be a similarity metric with non-trivial distribution-free right-tail and left-tail bounds u_R and u_L, and consider the test of Definition 7 with margin $t \geq 0$. Under H_0, $S^*_{d_X,d_Y}(\mathcal{D}_{XY}) \leq S^*_{d_{X'},d_{Y'}}(\mathcal{D}_{X'Y'}) + t$, for an observed test statistic $T_S = \Delta > t$ and $N, N' \in \mathbb{N}$, the tail probability $\mathbb{P}_{H_0}(T_S \geq \Delta)$ satisfies:*

$$\mathbb{P}_{H_0}(T_S \geq \Delta) \leq \inf_{0<\lambda<\Delta-t} \left[u_R(\lambda, N) + u_L(\Delta - t - \lambda, N')\right].$$

Moreover, if the samples (X,Y) and (X',Y') are independent, then

$$\mathbb{P}_{H_0}(T_S \geq \Delta) \leq \inf_{0<\lambda<\Delta-t} \left[u_R(\lambda, N) + u_L(\Delta - t - \lambda, N') - u_R(\lambda, N)u_L(\Delta - t - \lambda, N')\right].$$

See Appendix E.4 for the proof. Bounding the p-value from above provides a *distribution-free* criterion for rejecting the null hypothesis, offering principled significance claims without distributional assumptions. In practice, the infima in Theorem 2 are straightforward to obtain as for fixed N, N', Δ and t, the objective $F(\lambda) \coloneqq u_R(\lambda, N) + u_L(\Delta - t - \lambda, N')$ is a univariate function on the bounded interval $(0, \Delta - t)$, so the p-value bound for general samples can be efficiently approximated with a bounded scalar minimizer (Virtanen et al., 2020), which is how we evaluate all p-value bounds in our experiments. The independent-samples bound is evaluated the same way. We also derive further bounds for this independent-samples case in Corollary 3, but these are rarely tighter, and are thus omitted from the main experiments for brevity.

Comparing Empirical TSI and QSI Scores with a Medium-to-Large Gap Frequently Recovers the Population Ordering. To assess whether the p-value upper bounds of Theorem 2 are informative in practice, we instantiate them for TSI, QSI, and CKA-RBF using the tail bounds from Corollary 2 across varying sample sizes ($N = N'$) and excess margins $\Delta - t$ (Figure 4). In short, a low p-value upper bound guarantees that the observed empirical ordering reflects the true population ranking with high confidence. Owing to their sub-Gaussian decay, all three metrics exhibit a sharp phase transition where the bound drops from 1 to 0. Crucially, for TSI and QSI, this transition occurs at practical sample sizes ($N \in [10^3, 10^4]$) for moderate gaps ($\Delta - t \geq 0.05$). In contrast, for CKA-RBF the transition requires huge sample sizes ($N \geq 10^{10}$, even assuming an optimistic $m = 0.1$), rendering these bounds practically informative for TSI and QSI, but not for CKA-RBF.

5 Experiments

Following the evaluation protocol of Radford et al. (2021); Soares et al. (2026), we analyze multimodal vision-language alignment across vision-language models: CLIP (Radford et al., 2021),

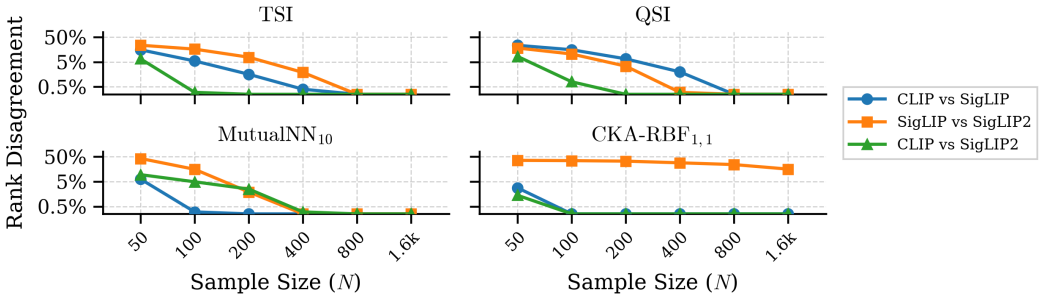

Figure 6: Finite-sample rankings disagree unpredictably with true population limits, showing disagreement rates that persist even at large sample sizes and diverge significantly across metrics and model comparisons.

SigLIP (Zhai et al., 2023), and SigLIP2 (Tschannen et al., 2025). Across all experiments, evaluations are performed on a fixed subset of 15k datapoints from the ImageNet validation set, pairing image representations with prompt-based label embeddings ("A photo of {label}").

5.1 Empirical Similarity Scores Can Misrepresent Population Similarity

We measure vision-language alignment in CLIP using Euclidean distance. To illustrate how empirical scores misrepresent alignment, we take the uniform empirical distribution over the aforementioned 15k fixed subset as our ground-truth data-generating process, using the similarity scores on this set as close approximations of its true population limits. We then evaluate each metric across 100 independent random subsets of size N drawn from this pool. Figure 5 shows that empirical scores computed on a limited batch often deviate from the population similarity of the data-generating process, and therefore misrepresent this quantity of interest. This result is noteworthy because existing work (Huh et al., 2024) uses batches of size 1,024, and the quadratic time complexity of these metrics makes it infeasible to use larger batches to close this gap.

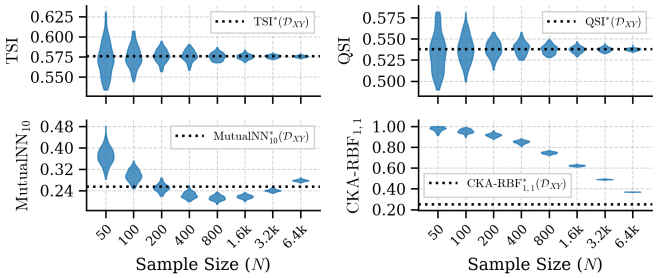

Figure 5: Empirical scores deviate from their population counterparts when measuring vision-language alignment in CLIP.

5.2 The Reliability of Empirical Scores Comparisons Depends on the Similarity Metric and Data Distribution Considered

We compare vision-language alignment across CLIP, SigLIP, and SigLIP2, measuring image-text alignment with TSI, QSI, MutualNN$_{10}$, and CKA-RBF$_{1,1}$ under cosine distance, matching their contrastive training objectives. For each model pair and sample size $N \in [50, 1.6\text{k}]$, we draw 1,000 independent subsets and define *rank disagreement* as the minimum proportion of runs favoring either model, which lower-bounds the disagreement rate with the true population ranking. Figure 6 illustrates the fundamental practical pitfall, where running repeated evaluations on independent finite subsets of the same size yields conflicting model rankings depending both on the chosen metric and the representations compared. Because finite sample estimates stabilize at unpredictable rates, a practitioner evaluating a single finite batch cannot know whether their observed ordering is consistent with the population rank. Furthermore, this behavior is not uniform across settings, as which model pair converges fastest changes with the metric, and convergence trajectories within the same metric also diverge markedly. Under MutualNN$_{10}$, the CLIP comparisons begin at similar disagreement rates before separating, while under CKA-RBF$_{1,1}$ the SigLIP versus SigLIP2 comparison suffers persistent instability, showing a 15.9% disagreement rate even at $N = 1.6\text{k}$. To resolve this instability, we next show how our hypothesis testing framework enables principled, statistically grounded comparisons directly from finite samples.

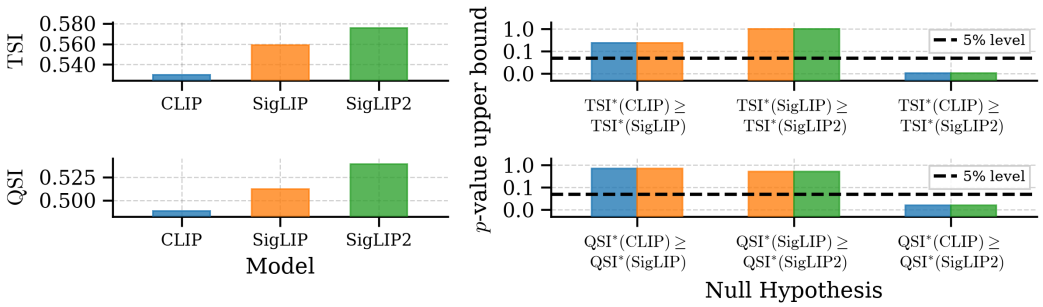

Figure 7: Image–text TSI and QSI increase from CLIP to SigLIP to SigLIP2 (left). At $N = 15\text{k}$ and $t = 0$, the p-value upper bounds confirm a population-level gap only for CLIP vs. SigLIP2 (right).

5.3 Statistically Grounded Conclusions at the Population Level - A Multimodal Similarity Case Study

We restrict this analysis to TSI and QSI, as the distribution-free tail bounds of CKA-RBF converge too slowly to be practically useful (Figure 4) and MutualNN admits no such bounds (Corollary 2). Every model is scored on the same paired sample of size $N = N' = 15\text{k}$, and we evaluate the test of Definition 7 with margin $t = 0$ using the general-sample p-value upper bound of Theorem 2. Reported ImageNet zero-shot accuracies increase from CLIP (68.3) to SigLIP (76.2) to SigLIP2 (78.2) (Tschannen et al., 2025). For each pair, the null hypothesis is the reverse of this ranking, and we reject it at the 5% level. In Figure 7 (left) we see that the empirical similarity between the image and text representations of CLIP, SigLIP and SigLIP2 is positively correlated with zero-shot performance, as TSI and QSI increase monotonically from CLIP to SigLIP to SigLIP2. Nevertheless, as motivated in the previous section, it is unclear whether these empirical differences are consistent with population-level differences. Here the p-value upper bounds of the statistical test play a critical role. In Figure 7 (right), only CLIP vs. SigLIP2 yields a p-value upper bound below 0.05. We therefore reject the null that CLIP is at least as similar as SigLIP2, and conclude that SigLIP2 is strictly more similar at the population level. The bounds for CLIP vs. SigLIP and SigLIP vs. SigLIP2 remain above 0.05. This illustrates the usefulness of our framework. SigLIP and SigLIP2 are known to show higher ImageNet zero-shot accuracy than CLIP, but that comparison only captures cross-modal alignment for the first nearest neighbor on this validation set. Our framework instead supports conclusions about the global disposition of these representations at the population level.

6 Conclusion

In this paper, we shift the focus of representational alignment analysis from sample dependent empirical scores to their population counterparts, which represent the true similarity obtained in the infinite sample limit. To make this shift practical, we establish distribution-free tail bounds that characterize worst case convergence across arbitrary distributions, and design a one sided hypothesis test that determines whether empirical rankings reliably reflect true population level orderings. While we demonstrate these theoretical contributions on TSI, QSI, CKA-RBF, and MutualNN, our framework is general and applies directly to other existing or future metrics, although minor scope limitations exist (see Appendix C).

Equipped with these theoretical tools, we establish several key practical insights. First, because MutualNN lacks distribution-free tail bounds, it can converge arbitrarily slowly, meaning empirical estimates cannot reliably represent true population alignment. In contrast, for TSI and QSI, sample sizes between 1,000 and 10,000 with observed gaps above 0.05 allow practitioners to conclude with statistical confidence that empirical rankings reflect population orderings. Our hypothesis testing framework directly translates these guarantees into practice, as demonstrated in our multimodal case study where we show that image and text representations in SigLIP2 are strictly more aligned at the population level than those in CLIP. Ultimately, our work demonstrates that empirical similarity scores alone can be deceptive, and establishing statistically grounded population limits is essential for drawing trustworthy scientific conclusions about representation learning.

AI USE STATEMENT

In this work, we used generative AI tools to provide critical ingredients for proving mathematical claims, assist in the writing of proofs. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, support qualitative and thematic data analysis, interpret results, and generate synthetic data sets, clean and reformat dataset are not applicable to this work. Additionally, we used generative AI tools to create or edit software code, edit a research paper to improve readability. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

LLMs were not used for ideation of corollaries/theorems/lemmas, however in certain instances LLMs were used to improve certain theoretical contributions. We give a more detailed report of our usage for transparency purposes:

1. LLMs were used to check the correctness of individual steps in the proofs. This check does not replace a formal proof assistant, and the authors remain responsible for the arguments.
2. LLMs were used to improve one of the reported bounds for Corollary 3, which is slightly tighter than the bound originally derived by the authors.
3. Although not reported in the paper, LLMs were used to construct a counter-example to show that not all MutualNN distances/distributions have non-increasing expectations. This counter-example served as an inspiration for the MutualNN proof in Appendix E.3, however that proof was human-derived and still deviates substantially from the aforementioned counter-example.
4. All other proofs were derived by the authors. LLMs were used only to help structure those proofs after they had been derived.

We believe that most theoretical proofs in this paper could have been produced by LLMs. We decided against using them from the outset, as the intuition gained while deriving the proofs helped us better design the proposed framework. We ultimately used them when we felt that doing so would enrich the manuscript.

REFERENCES

Gérard Biau and Luc Devroye. *Lectures on the nearest neighbor method*. Springer, 2015.

Laure Ciernik, Lorenz Linhardt, Marco Morik, Jonas Dippel, Simon Kornblith, and Lukas Muttenthaler. Objective drives the consistency of representational similarity across datasets. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 10920–10948. PMLR, 2025.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13:795–828, 2012.

Tianyu Cui, Yogesh Kumar, Pekka Marttinen, and Samuel Kaski. Deconfounded representation similarity for comparison of neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 19138–19151. Curran Associates, Inc., 2022.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of cka as a similarity measure in deep learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1556–1568, 2021.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pp. 63–77. Springer, 2005.

Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the platonic representation hypothesis: An aristotelian view. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Comput. Surv.*, 57(9):242:1–242:52, 2025.

A Koepke, Daniil Zverev, Shiry Ginosar, and Alexei A Efros. Back into plato's cave: examining cross-modal representational convergence at scale. *arXiv preprint arXiv:2604.18572*, 2026.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning of latent diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. In *9th International Conference on Learning Representations, ICLR*, 2021.

Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. *PLoS computational biology*, 10(4):e1003553, 2014.

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*, 2021.

Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 1976.

Isaac J Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938.

Diogo Soares, Pankhil Gawade, Andrea Dittadi, and Ewa Szczurek. Scalable and interpretable representation alignment with ordinal similarity. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*. PMLR, 2026.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Christopher J. Cueva, Erin Grant, Iris I. A. Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nathan Cloos, Nikolaus Kriegeskorte, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew Kyle Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment. *Transactions on Machine Learning Research*, 2025.

Michael Tschannen, Alexey A. Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier J. Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *CoRR*, abs/2502.14786, 2025.

Aad W Van der Vaart. *Asymptotic statistics*. Cambridge university press, 2000.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *The Thirteenth International Conference on Learning Representations, ICLR*, 2025.

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

APPENDIX

We organize the Appendix as follows:

- Appendix A: Additional assumptions, theoretical justification for the CND condition and boundedness of CKA-RBF.
- Appendix B: Additional independent-samples p-value bounds via the moment-generating function (Corollary 3) and a closed form under sub-Gaussian tails (Corollary 4).
- Appendix C: Discussion of theoretical and practical limitations.
- Appendix D: Mathematical notation, statistical tools, and background properties of HSIC.
- Appendix E: Complete proofs for all theorems and corollaries presented in the main text.
  - Appendix E.1: Proof of Theorem 1 (Existence of Population Similarity).
  - Appendix E.2: Proof of Corollary 1 (Existence of Population Similarities and Consistency for TSI, QSI, CKA-RBF, and MutualNN).
  - Appendix E.3: Proof of Corollary 2 (Distribution-Free Tail Bounds for TSI, QSI, and CKA-RBF, and Counterexample for MutualNN).
  - Appendix E.4: Proof of Theorem 2 (P-value Bounds from Distribution-Free Tail Bounds).
- Appendix F: Additional proofs for corollaries and lemmas in the appendix.
  - Appendix F.1: Proof of Corollary 3 (Moment Generating Function Bound).
  - Appendix F.2: Proof of Corollary 4 (Closed-Form P-value Bounds under Sub-Gaussian Tail Bounds).
  - Appendix F.3: Proof of Lemma 1 (Boundedness of CKA-RBF).
  - Appendix F.4: Proof of Lemma 2 (Boundedness of HSIC with Bounded Kernels).
  - Appendix F.5: Proof of Lemma 3 (Lower Bound Preservation for Empirical HSIC).

A ADDITIONAL ASSUMPTIONS

To ensure that empirical scores and asymptotic limits are well-defined across all representation spaces, we collect here the remaining formal regularity conditions referenced in the main text.

Assumption 2 (Supplementary Regularity Conditions)**.** For every admissible sample size N, joint distribution $\mathcal{D}_{XY}$, and metrics d_X, d_Y under consideration, we assume the following regularity conditions hold:

1. CKA-RBF$_{\sigma_X,\sigma_Y}$: $\sigma_X, \sigma_Y > 0$ are fixed constants, and the squared distance matrices D_X^2, D_Y^2 are conditionally negative definite (CND): $c^\top D_X^2 c \le 0$ and $c^\top D_Y^2 c \le 0$ for all $c \in \mathbb{R}^N$ with $c^\top \mathbf{1} = 0$, where $(D_X)_{ij} = d_X(x_i, x_j)$ and $(D_Y)_{ij} = d_Y(y_i, y_j)$.
2. MutualNN$_k$: $\mathbb{P}_{(X,Y)\sim\mathcal{D}_{XY}^3}\big(O_{d_X}^{\Delta}(x_1, x_2, x_3) = 0 \text{ or } O_{d_Y}^{\Delta}(y_1, y_2, y_3) = 0\big) = 0$.

We discuss the theoretical role and justification of each condition below.

Why Fix the Bandwidths? While practical studies often set bandwidths via sample heuristics like the median pairwise distance (Kornblith et al., 2019), estimating σ from the sample makes the kernel function random. Fixing $\sigma_X, \sigma_Y > 0$ a priori ensures a deterministic kernel, allowing us to apply the concentration bounds of Cortes et al. (2012) and preserve distribution-free guarantees.

Why is the CND Condition Necessary? When constructing an RBF kernel $k_\sigma(x, x') = \exp\big(-d(x, x')^2/2\sigma^2\big)$ from an arbitrary metric d, the resulting kernel Gram matrix is not guaranteed to be positive semi-definite (PSD). By Schoenberg's theorem (Theorem 9), d^2 being conditionally negative definite (CND) is necessary and sufficient for the induced RBF kernel to be PSD for all bandwidths $\sigma > 0$.

This PSD property connects kernel evaluations directly to an underlying inner-product geometry. Without CND, the centered Gram matrices K_X^c and K_Y^c can have negative eigenvalues, breaking this geometric structure and potentially yielding negative alignments $\mathrm{tr}(K_X^c K_Y^c) < 0$. Thus, CND is the exact condition ensuring that empirical CKA-RBF scores remain well-defined and strictly bounded in $[0, 1]$, as formalized below.

Lemma 1 (Boundedness of CKA-RBF)**.** *Under Assumption 2, empirical* CKA-RBF *satisfies*

$$\text{CKA-RBF}_{\sigma_X,\sigma_Y}(X, Y) \in [0, 1].$$

See proof in Appendix F.3.

Why Rule Out Distance Ties? When computing MutualNN, the sets $\text{KNN}_{k,d}$ are not guaranteed to be uniquely defined if multiple points lie at the same distance. Standard formulations resolve this with arbitrary tie-breaking conventions (Biau & Devroye, 2015). Condition 2 ensures that such ties occur with probability zero, so the empirical score is uniquely determined almost surely.

B Additional P-value Bounds from Distribution-Free Tail Bounds

Although omitted from the main text for brevity, we can derive alternative analytical bounds on the p-value for the test in Definition 7 when assuming independent samples. Rather than decomposing the tail event via a union bound (as in Appendix E.4), we apply Chernoff's bounding technique via the moment-generating function (MGF), establishing Corollary 3 (proof in Appendix F.1).

Corollary 3 (Additional P-value Bounds from Distribution-Free Tail Bounds)**.** *Let S be a bounded similarity metric in $[0, 1]$ with non-trivial distribution-free right-tail and left-tail bounds u_R and u_L, and consider the test of Definition 7 with margin $t \ge 0$. Under H_0, $S^*_{d_X,d_Y}(\mathcal{D}_{XY}) \le S^*_{d_{X'},d_{Y'}}(\mathcal{D}_{X'Y'}) + t$ with (X, Y) and (X', Y') independent, for an observed test statistic $T_S = \Delta > t$ and $N, N' \in \mathbb{N}$, the tail probability $\mathbb{P}_{H_0}(T_S \ge \Delta)$ satisfies:*

$$\mathbb{P}_{H_0}(T_S \ge \Delta) \le \inf_{\lambda \ge 0} \frac{\big(1 + \lambda \int_0^1 e^{\lambda s}\, u_R(s, N)\, ds\big)\big(1 + \lambda \int_0^1 e^{\lambda s}\, u_L(s, N')\, ds\big)}{e^{\lambda(\Delta - t)}}.$$

Analyzing the asymptotic properties of this variational bound is generally non-trivial. However, we can characterize its behavior for the sub-Gaussian tail bounds derived in this paper, which take the form

$$u_R(\varepsilon, N) = u_L(\varepsilon, N) = a \exp\big(-g(N)\varepsilon^2\big),$$

where $a > 0$ and g is a non-decreasing, approximately linear function of sample size. For such tails, setting $\lambda = \min\big\{\sqrt{g(N)}, \sqrt{g(N')}\big\}$ balances the exponential decay against the MGF integrals, yielding the explicit closed-form upper bound in Corollary 4.

Corollary 4 (Closed-Form P-value Bounds under Sub-Gaussian Tail Bounds)**.** *Let S be a bounded similarity metric in $[0, 1]$ with distribution-free tail bounds of the form*

$$u_R(\varepsilon, N) = u_L(\varepsilon, N) = a \exp\big(-g(N)\varepsilon^2\big),$$

where $a > 0$ and $g : \mathbb{N} \to \mathbb{R}_+$ satisfies $g(N) > 0$. Consider the test of Definition 7 with margin $t \geq 0$. Under H_0, with (X, Y) and (X', Y') independent, an observed test statistic $T_S = \Delta > t$, and $N, N' \in \mathbb{N}$,

$$\mathbb{P}_{H_0}(T_S \geq \Delta) \leq \left(1 + a\sqrt{\pi}\, e^{1/4}\right)^2 \exp\Big(-\min\big\{\sqrt{g(N)}, \sqrt{g(N')}\big\}(\Delta - t)\Big).$$

See Appendix F.2 for the complete proof. Asymptotically in N, this bound is looser than the union-bound formulation of Theorem 2, decaying with $\sqrt{g(N)}$ rather than $g(N)$. Moreover, practical application requires overcoming the prefactor $\left(1 + a\sqrt{\pi}e^{1/4}\right)^2 \approx 10.7$ (for $a = 1$). Nonetheless, this bound offers a distinct theoretical tradeoff as the exponent scales linearly in the excess margin $(\Delta - t)$ rather than quadratically, making it potentially advantageous in regimes with small sample sizes but substantial empirical gaps.

C Limitations

A few natural scope conditions of our setup warrant mention. We focus on bounded metrics and do not consider linear CKA, as extending these bounds to unbounded kernels would require strong moment restrictions that conflict with distribution-free guarantees. Additionally, while CKA-RBF admits exponential tail bounds, their practical convergence demands substantially larger sample sizes than ordinal metrics like TSI and QSI, which concentrate rapidly at accessible sample sizes. Our guarantees also rely on minimal regularity conditions necessary for formal rigor, including a priori fixed bandwidths and non-vanishing self-alignment to avoid degenerate geometries in CKA-RBF, as well as an intuitive expected-score monotonicity assumption for the asymptotic existence of MutualNN. Furthermore, our theoretical and empirical analyses only focus on four representative metrics, TSI, QSI, CKA-RBF, and MutualNN, serving as a blueprint that can be extended to broader metric families in future work. Finally, our tail bounds necessarily characterize worst-case convergence across all admissible distributions, and while specific empirical distributions may converge faster in practice, worst-case bounds provide the only mathematically sound guarantee one can assert from observed scores and sample sizes without unverifiable assumptions.

D Preliminaries and Notation

D.1 Notation

For clarity and conciseness, we adopt the following notation for expectations:

- $\mathbb{E}_{x,x',y,y'} := \mathbb{E}_{\substack{(x,y)\sim\mathcal{D}_{XY} \\ (x',y')\sim\mathcal{D}_{XY}}}$ denotes the expectation over two independent pairs (x, y) and (x', y') drawn identically from the joint distribution $\mathcal{D}_{XY}$.
- $\mathbb{E}_{x,y} := \mathbb{E}_{(x,y)\sim\mathcal{D}_{XY}}$ denotes the expectation over a single joint pair.
- $\mathbb{E}_x := \mathbb{E}_{x\sim\mathcal{D}_X}$ and $\mathbb{E}_{x,x'} := \mathbb{E}_{x,x'\sim\mathcal{D}_X}$ denote expectations over the marginal distribution $\mathcal{D}_X$ (with $\mathcal{D}_Y$ defined analogously).

D.2 General Asymptotic and Statistical Tools

Definition 8 (U-statistic (Van der Vaart, 2000))**.** Let $z_1, z_2, \ldots, z_N$ be independent and identically distributed random variables. A U-statistic of degree m with a symmetric kernel function $h(z_1, \ldots, z_m)$ is defined as:

$$U_N = \binom{N}{m}^{-1} \sum_{1\leq i_1<i_2<\cdots<i_m\leq N} h(z_{i_1}, z_{i_2}, \ldots, z_{i_m}).$$

Definition 9 (V-Statistic (Van der Vaart, 2000)). Let $z_1, \ldots, z_N$ be i.i.d. random variables taking values in a space $\mathcal{Z}$. A V-statistic of degree m with a symmetric kernel function $h : \mathcal{Z}^m \to \mathbb{R}$ is defined as:

$$V_N = \frac{1}{N^m} \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} \cdots \sum_{i_m=1}^{N} h(z_{i_1}, z_{i_2}, \ldots, z_{i_m}).$$

Theorem 3 (Law of Large Numbers for U-Statistics (Van der Vaart, 2000)). *Let $z_1, z_2, \ldots, z_N$ be independent and identically distributed random variables, and let $h(z_1, \ldots, z_m)$ be a symmetric kernel of degree $m \geq 1$ satisfying $\mathbb{E}[|h(z_1, \ldots, z_m)|] < \infty$.*

Then, as $N \to \infty$, U_N converges in probability, to its expectation:

$$U_N \xrightarrow[N\to\infty]{\mathbb{P}} \mathbb{E}[h(z_1, \ldots, z_m)].$$

Theorem 4 (Hoeffding's Inequality for U-Statistics (Hoeffding, 1963)). *If the kernel h is strictly bounded such that $a \leq h(\cdot) \leq b$ almost surely, then for any $\varepsilon > 0$*

$$\mathbb{P}\big(U_N - \mathbb{E}[U_N] \geq \varepsilon\big) \leq \exp\left(-\frac{2\lfloor N/m \rfloor \varepsilon^2}{(b-a)^2}\right),$$

and the two-sided inequality satisfies:

$$\mathbb{P}\big(|U_N - \mathbb{E}[U_N]| \geq \varepsilon\big) \leq 2\exp\left(-\frac{2\lfloor N/m \rfloor \varepsilon^2}{(b-a)^2}\right).$$

Note. Applying Theorem 4 to the reflected kernel $\tilde{h} = -h$, which satisfies $-b \leq \tilde{h}(\cdot) \leq -a$, immediately yields the corresponding lower-tail bound:

$$\mathbb{P}\big(\mathbb{E}[U_N] - U_N \geq \varepsilon\big) \leq \exp\left(-\frac{2\lfloor N/m \rfloor \varepsilon^2}{(b-a)^2}\right).$$

Theorem 5 (Continuous Mapping Theorem (Van der Vaart, 2000)). *If X_N converges in probability to θ and $g(\cdot)$ is a real-valued function continuous at the point θ, then $g(X_N)$ converges in probability to $g(\theta)$:*

$$X_N \xrightarrow[N\to\infty]{\mathbb{P}} \theta \quad \Rightarrow \quad g(X_N) \xrightarrow[N\to\infty]{\mathbb{P}} g(\theta)$$

Theorem 6 (Joint Convergence to Constants (Van der Vaart, 2000)). *Let X_N, Y_N, Z_N be sequences of random variables that each converge in probability to deterministic constants:*

$$X_N \xrightarrow[N\to\infty]{\mathbb{P}} x, \quad Y_N \xrightarrow[N\to\infty]{\mathbb{P}} y, \quad \text{and} \quad Z_N \xrightarrow[N\to\infty]{\mathbb{P}} z.$$

Then, their joint vector also converges in probability to the constant vector:

$$\begin{pmatrix} X_N \\ Y_N \\ Z_N \end{pmatrix} \xrightarrow[N\to\infty]{\mathbb{P}} \begin{pmatrix} x \\ y \\ z \end{pmatrix}.$$

Theorem 7 (Monotone Convergence Theorem for Real Sequences (Rudin, 1976)). *Let $(a_N)_{N=1}^{\infty}$ be a sequence of real numbers. If $(a_N)_{N=1}^{\infty}$ is bounded and monotone (either non-decreasing, $a_N \leq a_{N+1}$, or non-increasing, $a_N \geq a_{N+1}$ for all N), then it converges to a finite limit in $\mathbb{R}$.*

Theorem 8 (Cauchy-Schwarz Inequality (Rudin, 1976)). *Let $\mathcal{V}$ be an inner product space over $\mathbb{R}$ equipped with an inner product $\langle \cdot, \cdot \rangle$ and induced norm $\|u\| := \sqrt{\langle u, u \rangle}$. Then, for all $u, v \in \mathcal{V}$,*

$$|\langle u, v \rangle| \leq \|u\| \, \|v\|.$$

Theorem 9 (Schoenberg's Theorem (Schoenberg, 1938)). *The kernel $k_\sigma(x, x') = \exp(-d(x, x')^2/2\sigma^2)$ is positive semi-definite for all bandwidth parameters $\sigma > 0$ if and only if d^2 is conditionally negative definite (CND), i.e., for any finite sample $\{x_i\}_{i=1}^N \subset \mathcal{X}$ and its element-wise squared distance matrix D^2,*

$$c^\top D^2 c \leq 0 \quad \text{for all } c \in \mathbb{R}^N \text{ with } c^\top \mathbf{1} = 0.$$

D.3 Hilbert-Schmidt Independence Criterion (HSIC)

Definition 10 (Empirical Estimator of HSIC)**.** Given an i.i.d. sample $(X, Y) = \{(x_i, y_i)\}_{i=1}^N \sim \mathcal{D}_{XY}^N$, let $K_X, K_Y \in \mathbb{R}^{N\times N}$ denote the empirical Gram matrices with entries $(K_X)_{ij} = k_X(x_i, x_j)$ and $(K_Y)_{ij} = k_Y(y_i, y_j)$. The empirical estimator of HSIC is defined as:

$$\mathrm{HSIC}_{k_X,k_Y}(X, Y) = \frac{1}{(N-1)^2}\,\mathrm{tr}(K_X H K_Y H),$$

where $H = I_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top \in \mathbb{R}^{N\times N}$ is the centering matrix defined via the $N \times N$ identity matrix I_N and the all-ones vector $\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^N$.

Lemma 2 (Boundedness of HSIC)**.** *Let $k_X : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $k_Y : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be bounded kernels satisfying $|k_X(x, x')| \le c_x$ and $|k_Y(y, y')| \le c_y$ for all $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$. Then, the empirical estimator $\mathrm{HSIC}_{k_X,k_Y}(X, Y)$ is bounded.*

See proof in Appendix F.4.

Theorem 10 (Consistency of HSIC (Gretton et al., 2005))**.** *Let $k_X : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $k_Y : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be bounded kernels satisfying $|k_X(x, x')| \le c_x$ and $|k_Y(y, y')| \le c_y$ for all $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$. Then, for any joint probability distribution $\mathcal{D}_{XY}$, sample size $N \ge 2$, and confidence parameter $\delta \in (0, 1)$, with probability at least $1 - \delta$:*

$$\left|\mathrm{HSIC}_{k_X,k_Y}(X, Y) - \mathrm{HSIC}^*_{k_X,k_Y}(\mathcal{D}_{XY})\right| \le \sqrt{\frac{\ln(6/\delta)}{\alpha^2 N}} + \frac{C}{N},$$

where α and C are positive constants depending only on c_x and c_y. In particular, as $N \to \infty$, the right-hand side converges to zero, which implies consistency:

$$\mathrm{HSIC}_{k_X,k_Y}(X, Y) \xrightarrow[N\to\infty]{\mathbb{P}} \mathrm{HSIC}^*_{k_X,k_Y}(\mathcal{D}_{XY}).$$

Lemma 3 (Lower Bound Preservation for Empirical HSIC)**.** *Let k_X be a positive semi-definite kernel, and let $m \in \mathbb{R}$. Suppose that for every $N \ge 2$,*

$$\mathbb{P}\big(\mathrm{HSIC}_{k_X,k_X}(X, X) \ge m\big) = 1.$$

If $\mathrm{HSIC}_{k_X,k_X}(X, X)$ converges in probability to the deterministic constant $\mathrm{HSIC}^*_{k_X,k_X}(\mathcal{D}_{XX})$ as $N \to \infty$, then:

$$\mathrm{HSIC}^*_{k_X,k_X}(\mathcal{D}_{XX}) \ge m.$$

See proof in Appendix F.5.

E Proofs

E.1 Existence of Population Similarity

Theorem 1 (Existence of Population Similarity)**.** *Let S be a bounded similarity metric, i.e., $\mathrm{Im}(S) \subseteq [0, 1]$, and suppose that for every joint distribution $\mathcal{D}_{XY}$ and distance functions d_X, d_Y there exists $c \in [0, 1]$ such that $S_{d_X,d_Y}(X, Y) \xrightarrow[N\to\infty]{\mathbb{P}} c$ for $(X, Y) \sim \mathcal{D}_{XY}^N$. Then, for every $\mathcal{D}_{XY}, d_X, d_Y$, $S^*_{d_X,d_Y}(\mathcal{D}_{XY})$ exists in $[0, 1]$ and equals the in-probability limit of $S_{d_X,d_Y}(X, Y)$. In particular, S is consistent.*

Proof. Let $\mathcal{D}_{XY}$ be a fixed joint distribution, note that to show the entire theorem it suffices to show that:

$$\lim_{N\to\infty} \mathbb{E}_{(X,Y)\sim\mathcal{D}_{XY}^N}\left[S_{d_X,d_Y}(X, Y)\right] \overset{?}{=} c$$

This is because if such an identity is true, then the population similarity trivially exists and equals c by definition. Moreover, since by the corollary assumption $S_{d_X,d_Y}(X, Y)$ already converges in probability to c, if c equals the population similarity then $S_{d_X,d_Y}(X, Y)$ converges in probability to the population similarity and is therefore consistent.

To achieve this, fix $\varepsilon > 0$, and note the following chain of equalities, where the first derivation stems from the fact that $|S_{d_X,d_Y}(X,Y) - c| \leq 1$:

$$\begin{aligned}
&\mathbb{E}_{(X,Y)\sim\mathcal{D}_{XY}^N}\left[|S_{d_X,d_Y}(X,Y) - c|\right] \leq \\
&\mathbb{P}_{(X,Y)\sim\mathcal{D}_{XY}^N}\left(|S_{d_X,d_Y}(X,Y) - c| \geq \varepsilon\right) + \mathbb{P}_{(X,Y)\sim\mathcal{D}_{XY}^N}\left(|S_{d_X,d_Y}(X,Y) - c| < \varepsilon\right) \cdot \varepsilon \leq \\
&\mathbb{P}_{(X,Y)\sim\mathcal{D}_{XY}^N}\left(|S_{d_X,d_Y}(X,Y) - c| \geq \varepsilon\right) + \varepsilon
\end{aligned}$$

This implies that for all $\varepsilon > 0$:

$$0 \leq \mathbb{E}_{(X,Y)\sim\mathcal{D}_{XY}^N}\left[|S_{d_X,d_Y}(X,Y) - c|\right] \leq \mathbb{P}_{(X,Y)\sim\mathcal{D}_{XY}^N}\left(|S_{d_X,d_Y}(X,Y) - c| \geq \varepsilon\right) + \varepsilon$$

Taking the limit superior as $N \to \infty$ and using convergence in probability, we obtain for all $\varepsilon > 0$:

$$\begin{aligned}
0 &\leq \liminf_{N\to\infty} \mathbb{E}_{(X,Y)\sim\mathcal{D}_{XY}^N}\left[|S_{d_X,d_Y}(X,Y) - c|\right] \\
&\leq \limsup_{N\to\infty} \mathbb{E}_{(X,Y)\sim\mathcal{D}_{XY}^N}\left[|S_{d_X,d_Y}(X,Y) - c|\right] \\
&\leq \lim_{N\to\infty}\left(\mathbb{P}_{(X,Y)\sim\mathcal{D}_{XY}^N}\left(|S_{d_X,d_Y}(X,Y) - c| \geq \varepsilon\right) + \varepsilon\right) = \varepsilon.
\end{aligned}$$

Since this holds for every $\varepsilon > 0$,

$$\begin{aligned}
0 &\leq \liminf_{N\to\infty} \mathbb{E}_{(X,Y)\sim\mathcal{D}_{XY}^N}\left[|S_{d_X,d_Y}(X,Y) - c|\right] \\
&\leq \limsup_{N\to\infty} \mathbb{E}_{(X,Y)\sim\mathcal{D}_{XY}^N}\left[|S_{d_X,d_Y}(X,Y) - c|\right] \\
&\leq \inf_{\varepsilon>0} \varepsilon = 0,
\end{aligned}$$

and therefore

$$\begin{aligned}
&\lim_{N\to\infty} \mathbb{E}_{(X,Y)\sim\mathcal{D}_{XY}^N}\left[|S_{d_X,d_Y}(X,Y) - c|\right] = 0 \\
\implies &\lim_{N\to\infty} \mathbb{E}_{(X,Y)\sim\mathcal{D}_{XY}^N}\left[S_{d_X,d_Y}(X,Y)\right] = c.
\end{aligned}$$

□

E.2 Existence of Population Similarity and Consistency of Similarity Metrics

Corollary 1 (Existence of Population Similarity and Consistency of Similarity Metrics)**.** *Under Assumption 1, for every joint distribution $\mathcal{D}_{XY}, d_X, d_Y$,* TSI*,* QSI*, and* CKA-RBF$_{\sigma_X,\sigma_Y}$ *have population similarities and are consistent.* MutualNN$_k$ *has a population similarity if, for all sufficiently large N, $\mathbb{E}_{(X,Y)\sim\mathcal{D}_{XY}^N}[\text{MutualNN}_k(X,Y)] \geq \mathbb{E}_{(X,Y)\sim\mathcal{D}_{XY}^{N+1}}[\text{MutualNN}_k(X,Y)]$.*

E.2.1 Existence of Population Similarity and Consistency of TSI

We verify the two conditions of Theorem 1:

Boundedness: By Definition 1, $\text{TSI}_{d_X,d_Y}(X,Y)$ is defined as the empirical expectation of an indicator function over the set $\mathcal{T}$ of ordered triplets. Since indicators map to $\{0,1\}$, their empirical mean satisfies:

$$\text{Im}(\text{TSI}_{d_X,d_Y}) \subseteq [0,1],$$

which satisfies the boundedness condition of Theorem 1.

Convergence in Probability: Let $z_i = (x_i, y_i) \sim \mathcal{D}_{XY}$ for $i \in \{1,\ldots,N\}$. Notice that the set $\mathcal{T}$ contains all $N(N-1)(N-2)$ ordered permutations of 3 distinct items chosen from N samples. We can group these ordered triplets into their $\binom{N}{3}$ unique index subsets $\{i,j,k\}$ and sum over the $3! = 6$ permutations $\pi \in \mathcal{S}_3$ within each subset. This formulation allows us to rewrite $\text{TSI}_{d_X,d_Y}(X,Y)$ directly as a one-sample U-statistic of degree 3 (Definition 8):

$$\text{TSI}_{d_X,d_Y}(X,Y) = \binom{N}{3}^{-1} \sum_{1\leq i<j<k\leq N} h_{\text{TSI}}(z_i, z_j, z_k),$$

where h_{TSI} corresponds to the symmetric kernel obtained in Section N.9 of Soares et al. (2026):

$$h_{\mathrm{TSI}}(z_i, z_j, z_k) = \frac{1}{3!} \sum_{\pi \in \mathcal{S}_3} \mathbb{I}\left[O^{\Delta}_{d_X}(x_{\pi(i)}, x_{\pi(j)}, x_{\pi(k)}) = O^{\Delta}_{d_Y}(y_{\pi(i)}, y_{\pi(j)}, y_{\pi(k)})\right].$$

Because h_{TSI} is an average of indicator functions, it is bounded in $[0, 1]$, making the finite expectation condition $\mathbb{E}[|h_{\mathrm{TSI}}|] < \infty$ trivial. By the Law of Large Numbers for U-statistics (Theorem 3), the sample statistic converges in probability as $N \to \infty$:

$$\mathrm{TSI}_{d_X, d_Y}(X, Y) \xrightarrow[N\to\infty]{\mathbb{P}} c_{\mathrm{TSI}} = \mathbb{E}_{(X,Y)\sim\mathcal{D}^3_{XY}}\left[h_{\mathrm{TSI}}(z_i, z_j, z_k)\right] \in [0, 1].$$

Both requirements of Theorem 1 are satisfied: TSI is bounded in $[0, 1]$ and converges in probability to a constant c_{TSI}. Therefore, the population counterpart $\mathrm{TSI}^*_{d_X, d_Y}(\mathcal{D}_{XY})$ exists, belongs to $[0, 1]$, equals c_{TSI}, and the estimator is consistent.

E.2.2 Existence of Population Similarity and Consistency of QSI

Similarly to TSI, we verify the two conditions of Theorem 1:

Boundedness: By Definition 1, $\mathrm{QSI}_{d_X, d_Y}(X, Y)$ is defined as the empirical expectation of an indicator function over the set $\mathcal{Q}$ of ordered quadruplets. Since indicators map to $\{0, 1\}$, their empirical mean satisfies:

$$\mathrm{Im}(\mathrm{QSI}_{d_X, d_Y}) \subseteq [0, 1],$$

which satisfies the boundedness condition of Theorem 1.

Convergence in Probability: Let $z_i = (x_i, y_i) \sim \mathcal{D}_{XY}$ for $i \in \{1, \ldots, N\}$. The set $\mathcal{Q}$ contains all $N(N-1)(N-2)(N-3)$ ordered permutations of 4 distinct items chosen from N samples. Grouping these ordered quadruplets into their $\binom{N}{4}$ unique index subsets $\{i, j, k, l\}$ and summing over all $4! = 24$ permutations $\pi \in \mathcal{S}_4$ within each subset allows us to express $\mathrm{QSI}_{d_X, d_Y}(X, Y)$ directly as a one-sample U-statistic of degree 4 (Definition 8):

$$\mathrm{QSI}_{d_X, d_Y}(X, Y) = \binom{N}{4}^{-1} \sum_{1 \le i < j < k < l \le N} h_{\mathrm{QSI}}(z_i, z_j, z_k, z_l),$$

where h_{QSI} corresponds to the symmetric kernel obtained in Section N.9 of Soares et al. (2026):

$$h_{\mathrm{QSI}}(z_i, z_j, z_k, z_l) = \frac{1}{4!} \sum_{\pi \in \mathcal{S}_4} \mathbb{I}\left[O_{d_X}(x_{\pi(i)}, x_{\pi(j)}, x_{\pi(k)}, x_{\pi(l)}) = O_{d_Y}(y_{\pi(i)}, y_{\pi(j)}, y_{\pi(k)}, y_{\pi(l)})\right].$$

Because h_{QSI} is an average of indicator functions, it is bounded in $[0, 1]$, so the finite expectation condition $\mathbb{E}[|h_{\mathrm{QSI}}|] < \infty$ holds trivially. By the Law of Large Numbers for U-statistics (Theorem 3), the sample statistic converges in probability as $N \to \infty$:

$$\mathrm{QSI}_{d_X, d_Y}(X, Y) \xrightarrow[N\to\infty]{\mathbb{P}} c_{\mathrm{QSI}} = \mathbb{E}_{(X,Y)\sim\mathcal{D}^4_{XY}}\left[h_{\mathrm{QSI}}(z_i, z_j, z_k, z_l)\right] \in [0, 1].$$

Both requirements of Theorem 1 are satisfied: QSI is bounded in $[0, 1]$ and converges in probability to a constant c_{QSI}. Therefore, the population counterpart $\mathrm{QSI}^*_{d_X, d_Y}(\mathcal{D}_{XY})$ exists, belongs to $[0, 1]$, equals c_{QSI}, and the estimator is consistent.

E.2.3 Existence of Population Similarity and Consistency of CKA-RBF

We verify the conditions of Theorem 1:

Boundedness: By Lemma 1, $\mathrm{Im}(\text{CKA-RBF}_{\sigma_X, \sigma_Y}) \subseteq [0, 1]$, which satisfies the boundedness condition of Theorem 1.

Convergence in Probability: The estimator $\mathrm{HSIC}_{\sigma_X, \sigma_Y}(X, Y)$ is a special case of the general empirical formulation $\mathrm{HSIC}_{k_X, k_Y}(X, Y)$ (Definition 10) instantiated with RBF kernels, which are bounded by 1 ($c_x = c_y = 1$). Consequently, Lemma 2 guarantees that the empirical $\mathrm{HSIC}_{\sigma_X, \sigma_Y}$

estimators are bounded, and Theorem 10 ensures their consistency. Specifically, each estimator converges in probability to its deterministic population counterpart as $N \to \infty$:

$$\mathrm{HSIC}_{\sigma_X,\sigma_Y}(X,Y) \xrightarrow[N\to\infty]{\mathbb{P}} \mathrm{HSIC}^*_{\sigma_X,\sigma_Y}(\mathcal{D}_{XY}),$$

$$\mathrm{HSIC}_{\sigma_X,\sigma_X}(X,X) \xrightarrow[N\to\infty]{\mathbb{P}} \mathrm{HSIC}^*_{\sigma_X,\sigma_X}(\mathcal{D}_{XX}),$$

$$\mathrm{HSIC}_{\sigma_Y,\sigma_Y}(Y,Y) \xrightarrow[N\to\infty]{\mathbb{P}} \mathrm{HSIC}^*_{\sigma_Y,\sigma_Y}(\mathcal{D}_{YY}).$$

By Theorem 6, convergence in probability to deterministic constants implies joint convergence in probability:

$$\begin{pmatrix} \mathrm{HSIC}_{\sigma_X,\sigma_Y}(X,Y) \\ \mathrm{HSIC}_{\sigma_X,\sigma_X}(X,X) \\ \mathrm{HSIC}_{\sigma_Y,\sigma_Y}(Y,Y) \end{pmatrix} \xrightarrow[N\to\infty]{\mathbb{P}} \begin{pmatrix} \mathrm{HSIC}^*_{\sigma_X,\sigma_Y}(\mathcal{D}_{XY}) \\ \mathrm{HSIC}^*_{\sigma_X,\sigma_X}(\mathcal{D}_{XX}) \\ \mathrm{HSIC}^*_{\sigma_Y,\sigma_Y}(\mathcal{D}_{YY}) \end{pmatrix}.$$

Next, define the function $g : \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ by

$$g(u,v,w) = \frac{u}{\sqrt{vw}}.$$

The function $g(u,v,w) = \frac{u}{\sqrt{vw}}$ is continuous on its domain because it is a composition of continuous elementary operations (multiplication, square root, and division) where the denominator is non-zero (for $v, w > 0$). Under Assumption 1, the empirical self-alignments are almost surely bounded below by a strictly positive constant $m > 0$, which by Lemma 3 ensures that their limits also satisfy $\mathrm{HSIC}^*_{\sigma_X,\sigma_X}(\mathcal{D}_{XX}) \geq m > 0$ and $\mathrm{HSIC}^*_{\sigma_Y,\sigma_Y}(\mathcal{D}_{YY}) \geq m > 0$.

Applying the Continuous Mapping Theorem (Theorem 5) yields:

$$\begin{aligned} &\text{CKA-RBF}_{\sigma_X,\sigma_Y}(X,Y) \\ &= g\Big(\mathrm{HSIC}_{\sigma_X,\sigma_Y}(X,Y),\ \mathrm{HSIC}_{\sigma_X,\sigma_X}(X,X),\ \mathrm{HSIC}_{\sigma_Y,\sigma_Y}(Y,Y)\Big) \\ &\xrightarrow[N\to\infty]{\mathbb{P}}\ c_{\mathrm{CKA}} = \frac{\mathrm{HSIC}^*_{\sigma_X,\sigma_Y}(\mathcal{D}_{XY})}{\sqrt{\mathrm{HSIC}^*_{\sigma_X,\sigma_X}(\mathcal{D}_{XX})\,\mathrm{HSIC}^*_{\sigma_Y,\sigma_Y}(\mathcal{D}_{YY})}} \in [0,1]. \end{aligned}$$

Both requirements of Theorem 1 are satisfied: $\text{CKA-RBF}_{\sigma_X,\sigma_Y}$ is bounded in $[0,1]$ and converges in probability to the constant c_{CKA}. Therefore, the population similarity $\text{CKA-RBF}^*_{\sigma_X,\sigma_Y}(\mathcal{D}_{XY})$ exists, belongs to $[0,1]$, equals c_{CKA}, and the estimator is consistent.

E.2.4 Existence of Population Similarity for MutualNN

Proof. For each index i, define the shared-neighbor set

$$\mathrm{SharedNN}_{k,d_X,d_Y}(X,Y,i) \coloneqq \mathrm{KNN}_{k,d_X}(X,i) \cap \mathrm{KNN}_{k,d_Y}(Y,i),$$

so that, equivalently to Definition 3,

$$\mathrm{MutualNN}_{k,d_X,d_Y}(X,Y) = \frac{1}{N}\sum_{i=1}^{N} \frac{|\mathrm{SharedNN}_{k,d_X,d_Y}(X,Y,i)|}{k}.$$

By Definition 4, the population similarity $\mathrm{MutualNN}^*_{k,d_X,d_Y}(\mathcal{D}_{XY})$ exists if and only if the sequence of expected sample scores

$$a_N \coloneqq \mathbb{E}_{(X,Y)\sim\mathcal{D}^N_{XY}}\left[\mathrm{MutualNN}_{k,d_X,d_Y}(X,Y)\right]$$

converges to a finite limit as $N \to \infty$. To establish this convergence, we apply the Monotone Convergence Theorem for real sequences (Theorem 7) by verifying two properties of $(a_N)_{N=k+1}^{\infty}$:

Boundedness: For every $i \in \{1,\dots,N\}$, the cardinality of the shared-neighbor set is bounded by the neighborhood size:

$$0 \leq |\mathrm{SharedNN}_{k,d_X,d_Y}(X,Y,i)| \leq k.$$

Dividing by k and averaging across all N observations ensures that the empirical estimator is bounded in $[0, 1]$:

$$\mathrm{Im}(\mathrm{MutualNN}_{k,d_X,d_Y}) \subseteq [0, 1].$$

By the linearity and monotonicity of expectation, the sequence of expectations is likewise bounded: $0 \leq \mathbb{E}_{(X,Y)\sim\mathcal{D}_{XY}^N}[\mathrm{MutualNN}_{k,d_X,d_Y}(X, Y)] \leq 1$ for all $N \geq k + 1$.

Monotonicity: By the corollary assumption, the expected similarity is eventually non-increasing in the sample size:

$$\mathbb{E}_{(X,Y)\sim\mathcal{D}_{XY}^N}[\mathrm{MutualNN}_{k,d_X,d_Y}(X, Y)] \geq \mathbb{E}_{(X,Y)\sim\mathcal{D}_{XY}^{N+1}}[\mathrm{MutualNN}_{k,d_X,d_Y}(X, Y)]$$

for all sufficiently large N.

Since the tail of the sequence of expectations is bounded and monotone, Theorem 7 guarantees that it converges to a finite limit $c_{\mathrm{MutualNN}} \in [0, 1]$. Therefore, the population similarity

$$\mathrm{MutualNN}^*_{k,d_X,d_Y}(\mathcal{D}_{XY}) = \lim_{N\to\infty} \mathbb{E}_{(X,Y)\sim\mathcal{D}_{XY}^N}[\mathrm{MutualNN}_{k,d_X,d_Y}(X, Y)]$$

exists and belongs to $[0, 1]$. □

E.3 Existence of Distribution-Free Tail Bounds

Corollary 2 (Existence of Distribution-Free Tail Bounds)**.** *Under Assumption 1,* TSI*,* QSI*, and* $\mathrm{CKA\text{-}RBF}_{\sigma_X,\sigma_Y}$ *admit the non-trivial distribution-free right- and left-tail bounds*

$$\begin{aligned}
\text{TSI:} \quad & u_R(\varepsilon, N) = u_L(\varepsilon, N) = \exp\big(-2\lfloor N/3\rfloor\varepsilon^2\big), \\
\text{QSI:} \quad & u_R(\varepsilon, N) = u_L(\varepsilon, N) = \exp\big(-2\lfloor N/4\rfloor\varepsilon^2\big), \\
\text{CKA-RBF:} \quad & u_R(\varepsilon, N) = u_L(\varepsilon, N) = \min\big(1,\, 6\exp\big(-m^2N\varepsilon^2/(64\cdot 18^2)\big)\big),
\end{aligned}$$

where $m > 0$ *is the constant from Assumption 1.* $\mathrm{MutualNN}_k$ *admits no such bounds.*

E.3.1 Bound for TSI

Proof. We provide an overview of the argument here; the complete proof and detailed derivation can be found in Section N.9 of Soares et al. (2026).

Let $(X, Y) \sim \mathcal{D}_{XY}^N$ be a finite dataset of N independent pairs $z_i = (x_i, y_i)$ sampled from the joint distribution $\mathcal{D}_{XY}$. As shown by Soares et al. (2026), $\mathrm{TSI}_{d_X,d_Y}(X, Y)$ is a U-statistic of degree $m = 3$ (Definition 8) with symmetric kernel $h_{\mathrm{TSI}}(z_i, z_j, z_k) \in [0, 1]$.

Because U-statistics are strictly unbiased for any sample size $N \geq m$, their finite-sample expectation equals the population similarity:

$$\mathbb{E}_{(X,Y)\sim\mathcal{D}_{XY}^N}[\mathrm{TSI}_{d_X,d_Y}(X, Y)] = \lim_{N\to\infty} \mathbb{E}_{(X,Y)\sim\mathcal{D}_{XY}^N}[\mathrm{TSI}_{d_X,d_Y}(X, Y)] = \mathrm{TSI}^*_{d_X,d_Y}(\mathcal{D}_{XY}).$$

Using this identity and applying Hoeffding's inequality for U-statistics (Theorem 4) with $a = 0$, $b = 1$, and $m = 3$ yields:

$$\mathbb{P}\Big(\mathrm{TSI}_{d_X,d_Y}(X, Y) - \mathrm{TSI}^*_{d_X,d_Y}(\mathcal{D}_{XY}) \geq \varepsilon\Big) \leq \exp\left(-2\lfloor\tfrac{N}{3}\rfloor\varepsilon^2\right)$$

and

$$\mathbb{P}\Big(\mathrm{TSI}^*_{d_X,d_Y}(\mathcal{D}_{XY}) - \mathrm{TSI}_{d_X,d_Y}(X, Y) \geq \varepsilon\Big) \leq \exp\left(-2\lfloor\tfrac{N}{3}\rfloor\varepsilon^2\right).$$

□

E.3.2 Bound for QSI

Proof. Similarly, as shown by Soares et al. (2026), $\mathrm{QSI}_{d_X,d_Y}(X, Y)$ is a U-statistic of degree $m = 4$ (Definition 8) with kernel $h_{\mathrm{QSI}}(z_i, z_j, z_k, z_l) \in [0, 1]$. Applying Theorem 4 with $m = 4$ and $b - a = 1$ gives:

$$\mathbb{P}\Big(\mathrm{QSI}_{d_X,d_Y}(X, Y) - \mathrm{QSI}^*_{d_X,d_Y}(\mathcal{D}_{XY}) \geq \varepsilon\Big) \leq \exp\left(-2\lfloor\tfrac{N}{4}\rfloor\varepsilon^2\right)$$

and

$$\mathbb{P}\Big(\mathrm{QSI}^*_{d_X,d_Y}(\mathcal{D}_{XY}) - \mathrm{QSI}_{d_X,d_Y}(X, Y) \geq \varepsilon\Big) \leq \exp\left(-2\lfloor\tfrac{N}{4}\rfloor\varepsilon^2\right).$$

□

E.3.3 BOUNDS FOR CKA

Proof. While the main text focuses on CKA-RBF (Definition 2), we establish this result for the general formulation of Centered Kernel Alignment (CKA) with arbitrary bounded kernels (Cortes et al., 2012; Kornblith et al., 2019), from which the statement for CKA-RBF follows as an immediate special case.

Specifically, let $\mathcal{D}_{XY}$ denote a joint probability distribution over $\mathcal{X} \times \mathcal{Y}$. Let $k_X : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $k_Y : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be positive-definite kernels bounded such that $|k_X(x, x')| \leq c_x$ and $|k_Y(y, y')| \leq c_y$ for all $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$. For an i.i.d. sample $(X, Y) \sim \mathcal{D}_{XY}^N$, let $K_X, K_Y \in \mathbb{R}^{N\times N}$ denote the corresponding Gram matrices with entries $(K_X)_{ij} = k_X(x_i, x_j)$ and $(K_Y)_{ij} = k_Y(y_i, y_j)$. General empirical CKA is defined as:

$$\mathrm{CKA}_{k_X,k_Y}(X, Y) = \frac{\mathrm{HSIC}_{k_X,k_Y}(X, Y)}{\sqrt{\mathrm{HSIC}_{k_X,k_X}(X, X)\,\mathrm{HSIC}_{k_Y,k_Y}(Y, Y)}},$$

where $\mathrm{HSIC}_{k_X,k_Y}(X, Y) = \frac{1}{(N-1)^2}\,\mathrm{tr}(K_X H K_Y H)$ with centering matrix $H = I_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$. In the specific case of CKA-RBF (Definition 2), $k_X(x, x') = \exp\big(-d_X(x, x')^2/(2\sigma_X^2)\big) \leq 1$ and $k_Y(y, y') = \exp\big(-d_Y(y, y')^2/(2\sigma_Y^2)\big) \leq 1$, yielding kernel bounds $c_x = c_y = 1$.

Following Cortes et al. (2012), define the centered kernel function k_X^c as:

$$k_X^c(x, x') = k_X(x, x') - \mathbb{E}_x[k_X(x, x')] - \mathbb{E}_{x'}[k_X(x, x')] + \mathbb{E}_{x,x'}[k_X(x, x')],$$

where the expectations follow the compact notation established in Appendix D.1. The k_Y^c is defined analogously. Let $K_X, K_Y \in \mathbb{R}^{N\times N}$ be the empirical Gram matrices with entries $(K_X)_{ij} = k_X(x_i, x_j)$ and $(K_Y)_{ij} = k_Y(y_i, y_j)$, and let $K_X^c = HK_XH$ and $K_Y^c = HK_YH$ be the centered Gram matrices, where $H = I_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top \in \mathbb{R}^{N\times N}$ is the centering matrix.

For two real matrices A, B, their Frobenius inner product is defined as $\langle A, B\rangle_F = \mathrm{tr}(A^\top B)$, which induces the Frobenius norm $\|A\|_F = \sqrt{\langle A, A\rangle_F}$. Finally, let β denote the scale constant:

$$\beta = \max\left(\frac{c_x c_y}{\mathbb{E}_{x,x'}[(k_X^c(x, x'))^2]}, \frac{c_x c_y}{\mathbb{E}_{y,y'}[(k_Y^c(y, y'))^2]}\right).$$

Under these definitions, Cortes et al. (2012) established that for any $\delta \in (0, 1]$:

$$\mathbb{P}\left(\left|\frac{\mathbb{E}_{x,x',y,y'}\left[k_X^c(x, x')k_Y^c(y, y')\right]}{\sqrt{\mathbb{E}_{x,x'}\left[(k_X^c(x, x'))^2\right]\mathbb{E}_{y,y'}\left[(k_Y^c(y, y'))^2\right]}} - \frac{\langle K_X^c, K_Y^c\rangle_F}{\|K_X^c\|_F\|K_Y^c\|_F}\right| \geq 18\beta\left[\frac{3}{N} + 8\sqrt{\frac{\ln(6/\delta)}{2N}}\right]\right) \leq \delta.$$

To simplify the deviation bound, our goal is to absorb the lower-order term $\frac{3}{N}$ into the dominant radical $\sqrt{\frac{\ln(6/\delta)}{2N}}$, specifically seeking an inequality of the form:

$$\frac{3}{N} + 8\sqrt{\frac{\ln(6/\delta)}{2N}} \leq (1 + \alpha) \cdot 8\sqrt{\frac{\ln(6/\delta)}{2N}}$$

for some constant $\alpha > 0$.

To establish this, observe that for all $N \geq 2$ and $\delta \in (0, 1]$,

$$\frac{9}{32N} \leq \frac{9}{64} < \alpha^2 \ln(6) \leq \alpha^2 \ln(6/\delta),$$

which holds whenever $\alpha > \frac{3}{8\sqrt{\ln(6)}} \approx 0.279$. Multiplying both sides by $\frac{32}{N}$ yields

$$\frac{9}{N^2} \leq \frac{32\alpha^2 \ln(6/\delta)}{N} = 64\alpha^2 \cdot \frac{\ln(6/\delta)}{2N}.$$

Taking the square root gives the desired bound on the linear term:

$$\frac{3}{N} \leq 8\alpha\sqrt{\frac{\ln(6/\delta)}{2N}}.$$

Adding $8\sqrt{\frac{\ln(6/\delta)}{2N}}$ to both sides and factoring the shared radical yields

$$\frac{3}{N} + 8\sqrt{\frac{\ln(6/\delta)}{2N}} \leq (1+\alpha) \cdot 8\sqrt{\frac{\ln(6/\delta)}{2N}}.$$

Multiplying across by 18β, the finite-sample deviation bound holds with probability at least $1-\delta$:

$$\mathbb{P}\left(\left|\frac{\mathbb{E}_{x,x',y,y'}\left[k_X^c(x,x')k_Y^c(y,y')\right]}{\sqrt{\mathbb{E}_{x,x'}\left[(k_X^c(x,x'))^2\right]\mathbb{E}_{y,y'}\left[(k_Y^c(y,y'))^2\right]}} - \frac{\langle K_X^c, K_Y^c\rangle_F}{\|K_X^c\|_F\|K_Y^c\|_F}\right| \geq (1+\alpha)\cdot 18\beta \cdot 8\sqrt{\frac{\ln\frac{6}{\delta}}{2N}}\right) \leq \delta.$$

Choosing $\alpha = \sqrt{2}-1$ (which satisfies $\alpha > \frac{3}{8\sqrt{\ln(6)}}$), setting the threshold

$$\varepsilon = (1+\alpha)\cdot 18\beta \cdot 8\sqrt{\frac{\ln(6/\delta)}{2N}},$$

and solving for δ yields the tail probability bound:

$$\mathbb{P}\left(\left|\frac{\mathbb{E}_{x,x',y,y'}\left[k_X^c(x,x')k_Y^c(y,y')\right]}{\sqrt{\mathbb{E}_{x,x'}\left[(k_X^c(x,x'))^2\right]\mathbb{E}_{y,y'}\left[(k_Y^c(y,y'))^2\right]}} - \frac{\langle K_X^c, K_Y^c\rangle_F}{\|K_X^c\|_F\|K_Y^c\|_F}\right| \geq \varepsilon\right) \leq 6\exp\left(-\frac{N\varepsilon^2}{64\cdot 18^2\cdot\beta^2}\right). \tag{1}$$

To establish the concentration of CKA, we compare these alignment quantities directly to CKA by proving the following four claims:

(a) **Population Equivalence**:

$$\mathrm{CKA}^*_{k_X,k_Y}(\mathcal{D}_{XY}) = \frac{\mathbb{E}_{x,x',y,y'}\left[k_X^c(x,x')k_Y^c(y,y')\right]}{\sqrt{\mathbb{E}_{x,x'}\left[\left(k_X^c(x,x')\right)^2\right]\cdot\mathbb{E}_{y,y'}\left[\left(k_Y^c(y,y')\right)^2\right]}},$$

(b) **Estimator Equivalence**:

$$\mathrm{CKA}_{k_X,k_Y}(X,Y) = \frac{\langle K_X^c, K_Y^c\rangle_F}{\|K_X^c\|_F\|K_Y^c\|_F},$$

(c) **Uniform Bounding of** β,

(d) **Final Concentration Bound and Non-Triviality**.

Proof of (a): Population Equivalence. To establish population equivalence, we proceed in two steps: first, we show that the large-sample limit of the expected value of the empirical estimator $\mathrm{CKA}(X,Y)$ converges to the ratio of population $\mathrm{HSIC}^*(\mathcal{D}_{XY})$ via the Continuous Mapping Theorem; second, we prove algebraically that this ratio matches the centered kernel alignment formulation of Cortes et al. (2012).

Step 1: By Definition 4, the population quantity is defined as:

$$\mathrm{CKA}^*_{k_X,k_Y}(\mathcal{D}_{XY}) = \lim_{N\to\infty}\mathbb{E}_{(X,Y)\sim\mathcal{D}_{XY}^N}\left[\mathrm{CKA}_{k_X,k_Y}(X,Y)\right].$$

Since the kernels are bounded, by Lemma 2 the empirical HSIC estimator is bounded and, by Theorem 10, consistent. Specifically, the individual estimators $\mathrm{HSIC}(X,Y)$, $\mathrm{HSIC}(X,X)$, and $\mathrm{HSIC}(Y,Y)$ each converge in probability to their deterministic population counterparts. By Theorem 6 convergence to constants implies joint convergence, so the corresponding three-dimensional random vector also converges in probability:

$$\begin{pmatrix}\mathrm{HSIC}_{k_X,k_Y}(X,Y)\\ \mathrm{HSIC}_{k_X,k_X}(X,X)\\ \mathrm{HSIC}_{k_Y,k_Y}(Y,Y)\end{pmatrix} \xrightarrow[N\to\infty]{\mathbb{P}} \begin{pmatrix}\mathrm{HSIC}^*_{k_X,k_Y}(\mathcal{D}_{XY})\\ \mathrm{HSIC}^*_{k_X,k_X}(\mathcal{D}_{XX})\\ \mathrm{HSIC}^*_{k_Y,k_Y}(\mathcal{D}_{YY})\end{pmatrix}.$$

The function $g(u, v, w) = \frac{u}{\sqrt{vw}}$ is continuous on its domain because it is a composition of continuous elementary operations (multiplication, square root, and division) where the denominator is non-zero (for $v, w > 0$). By Assumption 1 and Lemma 3: $\mathrm{HSIC}^*(\mathcal{D}_{XX}) \geq m > 0$ and $\mathrm{HSIC}^*(\mathcal{D}_{YY}) \geq m > 0$, so applying the Continuous Mapping Theorem (Theorem 5) yields:

$$\frac{\mathrm{HSIC}_{k_X,k_Y}(X,Y)}{\sqrt{\mathrm{HSIC}_{k_X,k_X}(X,X)\,\mathrm{HSIC}_{k_Y,k_Y}(Y,Y)}} \xrightarrow[N\to\infty]{\mathbb{P}} \frac{\mathrm{HSIC}^*_{k_X,k_Y}(\mathcal{D}_{XY})}{\sqrt{\mathrm{HSIC}^*_{k_X,k_X}(\mathcal{D}_{XX})\,\mathrm{HSIC}^*_{k_Y,k_Y}(\mathcal{D}_{YY})}}.$$

Consequently, by Theorem 1, because $\mathrm{CKA}(X,Y) \in [0,1]$, we have:

$$\mathrm{CKA}^*_{k_X,k_Y}(\mathcal{D}_{XY}) = \frac{\mathrm{HSIC}^*_{k_X,k_Y}(\mathcal{D}_{XY})}{\sqrt{\mathrm{HSIC}^*_{k_X,k_X}(\mathcal{D}_{XX})\,\mathrm{HSIC}^*_{k_Y,k_Y}(\mathcal{D}_{YY})}}.$$

Step 2: To establish equivalence with the centered alignment, it remains to show that:

$$\mathbb{E}_{x,x',y,y'}\left[k^c_X(x,x')k^c_Y(y,y')\right] = \mathrm{HSIC}^*_{k_X,k_Y}(\mathcal{D}_{XY}).$$

Showing this identity equates the numerators, and by setting $Y = X$, the normalizing denominators satisfy $\mathbb{E}_{x,x'}[(k^c_X(x,x'))^2] = \mathrm{HSIC}^*_{k_X,k_X}(\mathcal{D}_{XX})$, which directly equals the denominator expression used by Cortes et al. (2012).

Following Gretton et al. (2005), define the population HSIC as:

$$\begin{aligned}\mathrm{HSIC}^*_{k_X,k_Y}(\mathcal{D}_{XY}) = \mathbb{E}_{x,x',y,y'}[k_X(x,x')k_Y(y,y')] + \mathbb{E}_{x,x'}[k_X(x,x')]\mathbb{E}_{y,y'}[k_Y(y,y')] \\ - 2\,\mathbb{E}_{x,y}\left[\mathbb{E}_{x'}[k_X(x,x')]\mathbb{E}_{y'}[k_Y(y,y')]\right]\end{aligned}$$

where the expectations follow the compact notation established in Appendix D.1.

Expanding the product of the centered kernels $\mathbb{E}_{x,x',y,y'}[k^c_X(x,x')k^c_Y(y,y')]$ yields 16 terms

$$\begin{aligned}&\mathbb{E}_{x,x',y,y'}[k^c_X(x,x')k^c_Y(y,y')] = \\ &\quad\mathbb{E}_{x,x',y,y'}\Big[k_X(x,x')\cdot\Big(k_Y(y,y') - \mathbb{E}_y[k_Y(y,y')] - \mathbb{E}_{y'}[k_Y(y,y')] + \mathbb{E}_{y,y'}[k_Y(y,y')]\Big)\Big] \\ &- \mathbb{E}_{x,x',y,y'}\Big[\mathbb{E}_x[k_X(x,x')]\cdot\Big(k_Y(y,y') - \mathbb{E}_y[k_Y(y,y')] - \mathbb{E}_{y'}[k_Y(y,y')] + \mathbb{E}_{y,y'}[k_Y(y,y')]\Big)\Big] \\ &- \mathbb{E}_{x,x',y,y'}\Big[\mathbb{E}_{x'}[k_X(x,x')]\cdot\Big(k_Y(y,y') - \mathbb{E}_y[k_Y(y,y')] - \mathbb{E}_{y'}[k_Y(y,y')] + \mathbb{E}_{y,y'}[k_Y(y,y')]\Big)\Big] \\ &+ \mathbb{E}_{x,x',y,y'}\Big[\mathbb{E}_{x,x'}[k_X(x,x')]\cdot\Big(k_Y(y,y') - \mathbb{E}_y[k_Y(y,y')] - \mathbb{E}_{y'}[k_Y(y,y')] + \mathbb{E}_{y,y'}[k_Y(y,y')]\Big)\Big].\end{aligned}$$

Since scalar expectation terms such as $\mathbb{E}_{x,x'}[k_X(x,x')]$ and $\mathbb{E}_{y,y'}[k_Y(y,y')]$ are deterministic constants, they can be factored out of expectations. Notice that in the fourth row, the scalar constant $\mathbb{E}_{x,x'}[k_X(x,x')]$ factors out, multiplying $\mathbb{E}_{y,y'}[k^c_Y(y,y')] = 0$, so the entire last row cancels out. For the first three rows, distributing the final term $\mathbb{E}_{y,y'}[k_Y(y,y')]$ yields three expressions of the form $\mathbb{E}_{x,x'}[k_X(x,x')]\,\mathbb{E}_{y,y'}[k_Y(y,y')]$-one with a positive sign (from the first row) and two with negative signs (from the second and third rows). Summing these gives:

$$\begin{aligned}&\mathbb{E}_{x,x',y,y'}[k^c_X(x,x')k^c_Y(y,y')] = \\ &\qquad\mathbb{E}_{x,x',y,y'}\Big[k_X(x,x')\cdot\Big(k_Y(y,y') - \mathbb{E}_y[k_Y(y,y')] - \mathbb{E}_{y'}[k_Y(y,y')]\Big)\Big] \\ &\quad- \mathbb{E}_{x,x',y,y'}\Big[\mathbb{E}_x[k_X(x,x')]\cdot\Big(k_Y(y,y') - \mathbb{E}_y[k_Y(y,y')] - \mathbb{E}_{y'}[k_Y(y,y')]\Big)\Big] \\ &\quad- \mathbb{E}_{x,x',y,y'}\Big[\mathbb{E}_{x'}[k_X(x,x')]\cdot\Big(k_Y(y,y') - \mathbb{E}_y[k_Y(y,y')] - \mathbb{E}_{y'}[k_Y(y,y')]\Big)\Big] \\ &\quad- \mathbb{E}_{x,x'}[k_X(x,x')]\mathbb{E}_{y,y'}[k_Y(y,y')]\end{aligned}$$

By the independence of i.i.d. draws (x,y) and (x',y'), terms such as:

$$\mathbb{E}_{x,x',y,y'}\left[\mathbb{E}_x[k_X(x,x')]\mathbb{E}_{y'}[k_Y(y,y')]\right] = \mathbb{E}_{x,x'}[k_X(x,x')]\mathbb{E}_{y,y'}[k_Y(y,y')]$$

and

$$\mathbb{E}_{x,x',y,y'}\left[k_X(x,x')\mathbb{E}_{y'}[k_Y(y,y')]\right] = \mathbb{E}_{x,y}\left[\mathbb{E}_{x'}[k_X(x,x')]\cdot\mathbb{E}_{y'}[k_Y(y,y')]\right]$$

simplify directly to:

$$\mathbb{E}_{x,x',y,y'}\left[k_X^c(x,x')k_Y^c(y,y')\right] = \mathrm{HSIC}^*_{k_X,k_Y}(\mathcal{D}_{XY}).$$

Specializing to identical domains gives $\mathbb{E}_{x,x'}\left[(k_X^c(x,x'))^2\right] = \mathrm{HSIC}^*_{k_X,k_X}(\mathcal{D}_{XX})$ and $\mathbb{E}_{y,y'}\left[(k_Y^c(y,y'))^2\right] = \mathrm{HSIC}^*_{k_Y,k_Y}(\mathcal{D}_{YY})$. Therefore:

$$\frac{\mathbb{E}_{x,x',y,y'}\left[k_X^c(x,x')k_Y^c(y,y')\right]}{\sqrt{\mathbb{E}_{x,x'}\left[(k_X^c(x,x'))^2\right]\mathbb{E}_{y,y'}\left[(k_Y^c(y,y'))^2\right]}} = \frac{\mathrm{HSIC}^*_{k_X,k_Y}(\mathcal{D}_{XY})}{\sqrt{\mathrm{HSIC}^*_{k_X,k_X}(\mathcal{D}_{XX})\,\mathrm{HSIC}^*_{k_Y,k_Y}(\mathcal{D}_{YY})}}$$
$$= \mathrm{CKA}^*_{k_X,k_Y}(\mathcal{D}_{XY}).$$

Proof of (b): Estimator Equivalence. Let $H = I_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$ denote the centering matrix, and let $K_X^c = HK_XH$ be the centered Gram matrix. Using the Frobenius inner product $\langle K, L\rangle_F = \mathrm{tr}[K^\top L]$ (which simplifies to $\mathrm{tr}[KL]$ for symmetric matrices) and the fact that $H = H^2$, we have:

$$\langle K_X^c, K_Y^c\rangle_F = \langle HK_XH, HK_YH\rangle_F = \mathrm{tr}[HK_XHHK_YH] = \mathrm{tr}[K_XHK_YH].$$

The empirical HSIC estimator satisfies:

$$\mathrm{HSIC}_{k_X,k_Y}(X,Y) = \frac{1}{(N-1)^2}\mathrm{tr}[K_XHK_YH] = \frac{1}{(N-1)^2}\langle K_X^c, K_Y^c\rangle_F.$$

Substituting this into the definition of empirical CKA yields:

$$\mathrm{CKA}_{k_X,k_Y}(X,Y) = \frac{\langle K_X^c, K_Y^c\rangle_F}{\|K_X^c\|_F\|K_Y^c\|_F}.$$

Thus, the finite-sample concentration bound applies directly to CKA.

Proof of (c): Bounding β. By Assumption 1, for all $N \geq 2$:

$$\mathbb{P}_{(X,Y)\sim\mathcal{D}_{XY}^N}\left(\mathrm{HSIC}_{k_X,k_X}(X,X) \geq m \text{ and } \mathrm{HSIC}_{k_Y,k_Y}(Y,Y) \geq m\right) = 1.$$

Because the empirical estimator is consistent (Theorem 10), $\mathrm{HSIC}_{k_X,k_X}(X,X) \xrightarrow{\mathbb{P}} \mathrm{HSIC}^*_{k_X,k_X}(\mathcal{D}_{XX})$ as $N \to \infty$. By Lemma 3:

$$\mathrm{HSIC}^*_{k_X,k_X}(\mathcal{D}_{XX}) \geq m \quad \text{and} \quad \mathrm{HSIC}^*_{k_Y,k_Y}(\mathcal{D}_{YY}) \geq m.$$

Using the equivalence $\mathbb{E}_{x,x'}[(k_X^c(x,x'))^2] = \mathrm{HSIC}^*_{k_X,k_X}(\mathcal{D}_{XX}) \geq m$, the scaling parameter β is bounded uniformly by:

$$\beta = \max\left(\frac{c_xc_y}{\mathrm{HSIC}^*_{k_X,k_X}(\mathcal{D}_{XX})}, \frac{c_xc_y}{\mathrm{HSIC}^*_{k_Y,k_Y}(\mathcal{D}_{YY})}\right) \leq \frac{c_xc_y}{m}.$$

Squaring this inequality yields $\frac{1}{\beta^2} \geq \frac{m^2}{(c_xc_y)^2}$.

Proof of (d): Final Concentration Bound and Non-Triviality. Substituting the population equivalence from Step 1, the empirical equivalence from Step 2, and the uniform upper bound on β from Step 3 into equation 1 yields:

$$\mathbb{P}\left(\left|\mathrm{CKA}_{k_X,k_Y}(X,Y) - \mathrm{CKA}^*_{k_X,k_Y}(\mathcal{D}_{XY})\right| \geq \varepsilon\right) \leq 6\exp\left(-\frac{m^2}{64\cdot(18\cdot c_x\cdot c_y)^2}N\varepsilon^2\right).$$

Since the absolute deviation satisfies:

$$\mathbb{P}(A - B \geq \varepsilon) \leq \mathbb{P}(|A-B| \geq \varepsilon) \quad \text{and} \quad \mathbb{P}(B - A \geq \varepsilon) \leq \mathbb{P}(|A-B| \geq \varepsilon),$$

both the right-tail and left-tail bounds satisfy:

$$u_R(\varepsilon, N) = u_L(\varepsilon, N) = 6\exp\left(-\frac{m^2}{64\cdot(18\cdot c_x\cdot c_y)^2}N\varepsilon^2\right).$$

This produces a non-trivial tail bound that is strictly decreasing in both N and ε for all $N \geq 2$ and $\varepsilon > 0$, and decays exponentially to 0 as $N \to \infty$, which completes the proof.

In particular, for CKA-RBF the kernels satisfy $k(z, z') \in (0, 1]$, giving $c_x = c_y = 1$. Truncating at 1 to satisfy $\mathrm{Im}(u) \subseteq [0, 1]$, the bound simplifies to:

$$u_R(\varepsilon, N) = u_L(\varepsilon, N) = \min\left(1,\ 6\exp\left(-\frac{m^2 N\varepsilon^2}{64 \cdot 18^2}\right)\right).$$

□

E.3.4 Bound for MutualNN

Proof. For each index i, define the shared-neighbor set

$$\mathrm{SharedNN}_{k,d_X,d_Y}(X, Y, i) \coloneqq \mathrm{KNN}_{k,d_X}(X, i) \cap \mathrm{KNN}_{k,d_Y}(Y, i),$$

so that, equivalently to Definition 3,

$$\mathrm{MutualNN}_{k,d_X,d_Y}(X, Y) = \frac{1}{N}\sum_{i=1}^{N} \frac{|\mathrm{SharedNN}_{k,d_X,d_Y}(X, Y, i)|}{k}.$$

We will proceed to show this statement: $\mathrm{MutualNN}_{k,d_X,d_Y}$ does not admit *non-trivial distribution-free right-tail* nor *left-tail bounds*. This proof consists of showing two statements (a) $\mathrm{MutualNN}_{k,d_X,d_Y}$ does not admit *non-trivial distribution-free right-tail bounds*, and (b) $\mathrm{MutualNN}_{k,d_X,d_Y}$ does not admit *non-trivial distribution-free left-tail bounds*.

Assume by contradiction that $\mathrm{MutualNN}_{k,d_X,d_Y}$ admits a *non-trivial distribution-free right-tail bound* (conversely *left-tail*), then by definition for any distribution $\mathcal{D}_{XY}$ and any distance function d_X, d_Y, for all $\varepsilon > 0$ and $N \in \mathbb{N}$, we have that for u_R:

$$\mathbb{P}_{(X,Y)\sim\mathcal{D}_{XY}^N}\left(\mathrm{MutualNN}_{k,d_X,d_Y}(X, Y) - \mathrm{MutualNN}^*_{k,d_X,d_Y}(\mathcal{D}_{XY}) \geq \varepsilon\right) \leq u_R(\varepsilon, N),$$

conversely for u_L:

$$\mathbb{P}_{(X,Y)\sim\mathcal{D}_{XY}^N}\left(\mathrm{MutualNN}^*_{k,d_X,d_Y}(\mathcal{D}_{XY}) - \mathrm{MutualNN}_{k,d_X,d_Y}(X, Y) \geq \varepsilon\right) \leq u_L(\varepsilon, N)$$

where u_R (conversely, u_L) is a non-trivial distribution-free right-tail bound (conversely, left-tail bound), i.e., non-increasing in each argument and, for every $\varepsilon > 0$, $\lim_{N\to\infty} u_R(\varepsilon, N) = 0$ (conversely $\lim_{N\to\infty} u_L(\varepsilon, N) = 0$).

As a consequence of being non-trivial, if we fix $\varepsilon = 1$ and $1 > \Delta > 0$, then there exist (possibly different) $n_{0,R}, n_{0,L} \in \mathbb{N} \setminus \{1, \cdots, 2k+1\}$ such that $u_R(1, n_{0,R}) < \Delta$ and $u_L(1, n_{0,L}) < \Delta$, respectively. Therefore, for the right-tail statement it suffices to construct a counter-example such that:

$$\Delta < \mathbb{P}_{(X,Y)\sim\mathcal{D}_{XY}^{n_{0,R}}}\left(\mathrm{MutualNN}_{k,d_X,d_Y}(X, Y) - \mathrm{MutualNN}^*_{k,d_X,d_Y}(\mathcal{D}_{XY}) \geq 1\right) \leq u_R(1, n_{0,R})$$

Independently, for the left-tail statement it suffices to construct a counter-example such that:

$$\Delta < \mathbb{P}_{(X,Y)\sim\mathcal{D}_{XY}^{n_{0,L}}}\left(\mathrm{MutualNN}^*_{k,d_X,d_Y}(\mathcal{D}_{XY}) - \mathrm{MutualNN}_{k,d_X,d_Y}(X, Y) \geq 1\right) \leq u_L(1, n_{0,L})$$

Each inequality yields a contradiction for its corresponding tail. To present the two constructions uniformly, let n_0 denote $n_{0,R}$ in the right-tail case and $n_{0,L}$ in the left-tail case. We will construct counter-examples that take a parameter $\beta \in (0, 1)$ and state that for $\varepsilon = 1$:

$$\mathbb{P}_{(X,Y)\sim\mathcal{D}_{XY}^{n_0}}\left(\mathrm{MutualNN}_{k,d_X,d_Y}(X, Y) - \mathrm{MutualNN}^*_{k,d_X,d_Y}(\mathcal{D}_{XY}) \geq 1\right) = (1-\beta)^{n_0}$$

and

$$\mathbb{P}_{(X,Y)\sim\mathcal{D}_{XY}^{n_0}}\left(\mathrm{MutualNN}^*_{k,d_X,d_Y}(\mathcal{D}_{XY}) - \mathrm{MutualNN}_{k,d_X,d_Y}(X, Y) \geq 1\right) = (1-\beta)^{n_0}.$$

Given that for any $n_0 \in \mathbb{N} \setminus \{1, \cdots, 2k+1\}, 1 > \Delta > 0$, the equation $(1-\beta)^{n_0} > \Delta$ has a non-empty solution set $0 < \beta < 1 - \Delta^{\frac{1}{n_0}}$, this implies that no such independent upper bound can exist for counter-examples with this parameter, which will show exist by construction.

To see this consider the following counter-example for u_R:

Let $1 > \beta > 0, \mathcal{D}_{XY} = (A, B) \times (A, B), (A, B) \sim \text{Unif}((0,1)^2)$, now select d_X, d_Y to be the following distance functions:

$$d_X((a,b),(a',b')) = \begin{cases} 0 & \text{if } (a,b) = (a',b'), \\ 8 + |b - b'| & \text{if } a \geq \beta \wedge a' \geq \beta \wedge (a,b) \neq (a',b'), \\ 5 + |b - b'| & \text{otherwise.} \end{cases}$$

$$d_Y((a,b),(a',b')) = \begin{cases} 0 & \text{if } (a,b) = (a',b'), \\ 5 + |b - b'| & \text{if } a \geq \beta \wedge a' \geq \beta \wedge (a,b) \neq (a',b'), \\ 8 - |b - b'| & \text{otherwise.} \end{cases}$$

The distance functions d_X and d_Y defined above are indeed proper metrics. Both satisfy the identity of indiscernibles ($d(x,x) = 0$) and symmetry ($d(x,y) = d(y,x)$) by construction. To verify the triangle inequality, consider any three points (x_i, x_j, x_k). If any two points coincide (e.g., $x_i = x_j$), the inequality is trivially satisfied. In the case where all points are distinct, we have $d_X(x_i, x_j) \in [5, 9]$, so for any pair, the sum $d_X(x_i, x_j) + d_X(x_i, x_k) \in [10, 18]$, which always exceeds the maximum possible value for $d_X(x_j, x_k)$. A similar argument applies for d_Y. Thus, the triangle inequality holds in all cases.

Moreover, with this setup $\text{MutualNN}_{k,d_X,d_Y}(X,Y)$ almost surely for $n_0 \geq 2k+1$ depends exclusively on $c(X) \coloneqq |\{x_i | (x_i)_0 < \beta, i \in \{1, \cdots, n_0\}\}|$ such that :

$$\begin{aligned}
&\text{MutualNN}_{k,d_X,d_Y}(X,Y) \\
&= \frac{1}{n_0} \sum_{i=1}^{n_0} \frac{|\text{SharedNN}_{k,d_X,d_Y}(X,Y,i)|}{k} \\
&= \frac{1}{n_0} \left(\left(\sum_{\substack{i \text{ s.t.} \\ (x_i)_0 < \beta}} \frac{|\text{SharedNN}_{k,d_X,d_Y}(X,Y,i)|}{k} \right) + \left(\sum_{\substack{i \text{ s.t.} \\ (x_i)_0 \geq \beta}} \frac{|\text{SharedNN}_{k,d_X,d_Y}(X,Y,i)|}{k} \right) \right) \\
&= \frac{1}{n_0} \sum_{\substack{i \text{ s.t.} \\ (x_i)_0 \geq \beta}} \frac{\max(k - c(X), 0)}{k} \\
&= \frac{1}{n_0} (n_0 - c(X)) \frac{\max(k - c(X), 0)}{k} \\
&= \left(1 - \tfrac{c(X)}{n_0}\right) \max\left(1 - \tfrac{c(X)}{k}, 0\right)
\end{aligned}$$

As a consequence, we can derive the following chain of equalities:

$$\begin{aligned}
&\mathbb{E}_{(X,Y) \sim \mathcal{D}_{XY}^{n_0}} \left[\text{MutualNN}_{k,d_X,d_Y}(X,Y)\right] \\
&= \sum_{j=0}^{n_0} \mathbb{P}_{(X,Y) \sim \mathcal{D}_{XY}^{n_0}}(c(X) = j) \cdot \left(1 - \tfrac{j}{n_0}\right) \max\left(1 - \tfrac{j}{k}, 0\right) \\
&= \sum_{j=0}^{k-1} \mathbb{P}_{(X,Y) \sim \mathcal{D}_{XY}^{n_0}}(c(X) = j) \cdot \left(1 - \tfrac{j}{n_0}\right) \left(1 - \tfrac{j}{k}\right) \\
&= \sum_{j=0}^{k-1} \binom{n_0}{j} (1-\beta)^{n_0 - j} \cdot \beta^j \cdot \left(1 - \tfrac{j}{n_0}\right) \left(1 - \tfrac{j}{k}\right)
\end{aligned}$$

Note that this implies that:

$$\begin{aligned}
&\mathrm{MutualNN}^*_{k,d_X,d_Y}(\mathcal{D}_{XY}) \\
&= \lim_{N\to\infty} \mathbb{E}_{(X,Y)\sim\mathcal{D}^N_{XY}}\left[\mathrm{MutualNN}_{k,d_X,d_Y}(X,Y)\right] \\
&= \lim_{N\to\infty} \sum_{j=0}^{k-1} \binom{N}{j} (1-\beta)^{N-j}\cdot\beta^j\cdot\left(1-\tfrac{j}{N}\right)\left(1-\tfrac{j}{k}\right) \\
&= 0
\end{aligned}$$

This implies that for $\varepsilon = 1$:

$$\begin{aligned}
&\mathbb{P}_{(X,Y)\sim\mathcal{D}^{n_0}_{XY}}\left(\mathrm{MutualNN}_{k,d_X,d_Y}(X,Y) - \mathrm{MutualNN}^*_{k,d_X,d_Y}(\mathcal{D}_{XY}) \geq 1\right) \\
&= \mathbb{P}_{(X,Y)\sim\mathcal{D}^{n_0}_{XY}}\left(\mathrm{MutualNN}_{k,d_X,d_Y}(X,Y) \geq 1\right) \\
&= \mathbb{P}_{(X,Y)\sim\mathcal{D}^{n_0}_{XY}}\left(\mathrm{MutualNN}_{k,d_X,d_Y}(X,Y) = 1\right) \\
&= \mathbb{P}_{(X,Y)\sim\mathcal{D}^{n_0}_{XY}}\left(c(X) = 0\right) \\
&= (1-\beta)^{n_0}
\end{aligned}$$

And the following counter-example for u_L:

Let $1 > \beta > 0, \mathcal{D}_{XY} = (A,B)\times(A,B), (A,B) \sim \mathrm{Unif}((0,1)^2)$, now select d_X, d_Y to be the following distance functions:

$$d_X((a,b),(a',b')) = \begin{cases} 0, & \text{if } (a,b) = (a',b') \\ 8 + |b-b'|, & \text{if } a \geq \beta \wedge a' \geq \beta \wedge (a,b) \neq (a',b') \\ 5 + |b-b'|, & \text{otherwise} \end{cases}$$

$$d_Y((a,b),(a',b')) = \begin{cases} 0, & \text{if } (a,b) = (a',b') \\ 8 - |b-b'|, & \text{if } a \geq \beta \wedge a' \geq \beta \wedge (a,b) \neq (a',b') \\ 5 + |b-b'|, & \text{otherwise} \end{cases}$$

The distance functions d_X and d_Y defined above are indeed proper metrics. Both satisfy the identity of indiscernibles ($d(x,x) = 0$) and symmetry ($d(x,y) = d(y,x)$) by construction. To verify the triangle inequality, consider any three points (x_i, x_j, x_k). If any two points coincide (e.g., $x_i = x_j$), the inequality is trivially satisfied. In the case where all points are distinct, we have $d_X(x_i,x_j) \in [5,9]$, so for any pair, the sum $d_X(x_i,x_j) + d_X(x_i,x_k) \in [10,18]$, which always exceeds the maximum possible value for $d_X(x_j,x_k)$. A similar argument applies for d_Y. Thus, the triangle inequality holds in all cases.

Moreover, with this setup $\mathrm{MutualNN}_{k,d_X,d_Y}(X,Y)$ almost surely for $n_0 \geq 2k+1$ depends exclusively on $c(X) \coloneqq |\{x_i | (x_i)_0 < \beta, i \in \{1,\cdots,n_0\}\}|$ such that :

$$\begin{aligned}
&\mathrm{MutualNN}_{k,d_X,d_Y}(X,Y) \\
&= \frac{1}{n_0}\sum_{i=1}^{n_0} \frac{|\mathrm{SharedNN}_{k,d_X,d_Y}(X,Y,i)|}{k} \\
&= \frac{1}{n_0}\left(\left(\sum_{\substack{i \text{ s.t.} \\ (x_i)_0 < \beta}} \frac{|\mathrm{SharedNN}_{k,d_X,d_Y}(X,Y,i)|}{k}\right) + \left(\sum_{\substack{i \text{ s.t.} \\ (x_i)_0 \geq \beta}} \frac{|\mathrm{SharedNN}_{k,d_X,d_Y}(X,Y,i)|}{k}\right)\right) \\
&= \frac{c(X)}{n_0} + \frac{1}{n_0}\sum_{\substack{i \text{ s.t.} \\ (x_i)_0 \geq \beta}} \frac{\min(c(X),k)}{k} \\
&= \frac{c(X)}{n_0} + \left(1 - \frac{c(X)}{n_0}\right)\left(\frac{\min(c(X),k)}{k}\right)
\end{aligned}$$

As a consequence, we can derive the following chain of equalities:

$$
\begin{aligned}
&\mathbb{E}_{(X,Y)\sim\mathcal{D}_{XY}^{n_0}}\left[\mathrm{MutualNN}_{k,d_X,d_Y}(X,Y)\right] \\
&= \sum_{j=0}^{n_0} \mathbb{P}_{(X,Y)\sim\mathcal{D}_{XY}^{n_0}}(c(X)=j)\cdot\left(\tfrac{j}{n_0}+\left(1-\tfrac{j}{n_0}\right)\left(\tfrac{\min(j,k)}{k}\right)\right) \\
&= \sum_{j=0}^{n_0} \binom{n_0}{j}(1-\beta)^{n_0-j}\cdot\beta^j\cdot\left(\tfrac{j}{n_0}+\left(1-\tfrac{j}{n_0}\right)\left(\tfrac{\min(j,k)}{k}\right)\right) \\
&= \left(\sum_{j=k}^{n_0} \binom{n_0}{j}(1-\beta)^{n_0-j}\cdot\beta^j\right) \\
&\quad + \left(\sum_{j=0}^{k-1} \binom{n_0}{j}(1-\beta)^{n_0-j}\cdot\beta^j\cdot\left(\tfrac{j}{n_0}+\left(1-\tfrac{j}{n_0}\right)\left(\tfrac{\min(j,k)}{k}\right)\right)\right) \\
&= \left(\sum_{j=0}^{n_0} \binom{n_0}{j}(1-\beta)^{n_0-j}\cdot\beta^j\right) \\
&\quad - \left(\sum_{j=0}^{k-1} \binom{n_0}{j}(1-\beta)^{n_0-j}\cdot\beta^j\cdot\left(1-\left(\tfrac{j}{n_0}+\left(1-\tfrac{j}{n_0}\right)\left(\tfrac{\min(j,k)}{k}\right)\right)\right)\right) \\
&= 1-\sum_{j=0}^{k-1} \binom{n_0}{j}(1-\beta)^{n_0-j}\cdot\beta^j\cdot\left(1-\left(\tfrac{j}{n_0}+\left(1-\tfrac{j}{n_0}\right)\left(\tfrac{\min(j,k)}{k}\right)\right)\right)
\end{aligned}
$$

Note that this implies that:

$$
\begin{aligned}
&\mathrm{MutualNN}^*_{k,d_X,d_Y}(\mathcal{D}_{XY}) \\
&= \lim_{N\to\infty} \mathbb{E}_{(X,Y)\sim\mathcal{D}_{XY}^{N}}\left[\mathrm{MutualNN}_{k,d_X,d_Y}(X,Y)\right] \\
&= \lim_{N\to\infty} 1-\sum_{j=0}^{k-1} \binom{N}{j}(1-\beta)^{N-j}\cdot\beta^j\cdot\left(1-\left(\tfrac{j}{N}+\left(1-\tfrac{j}{N}\right)\left(\tfrac{\min(j,k)}{k}\right)\right)\right) \\
&= 1
\end{aligned}
$$

This implies that for $\varepsilon = 1$:

$$
\begin{aligned}
&\mathbb{P}_{(X,Y)\sim\mathcal{D}_{XY}^{n_0}}\left(\mathrm{MutualNN}^*_{k,d_X,d_Y}(\mathcal{D}_{XY}) - \mathrm{MutualNN}_{k,d_X,d_Y}(X,Y) \geq 1\right) \\
&= \mathbb{P}_{(X,Y)\sim\mathcal{D}_{XY}^{n_0}}\left(1-\mathrm{MutualNN}_{k,d_X,d_Y}(X,Y) \geq 1\right) \\
&= \mathbb{P}_{(X,Y)\sim\mathcal{D}_{XY}^{n_0}}\left(\mathrm{MutualNN}_{k,d_X,d_Y}(X,Y) \leq 0\right) \\
&= \mathbb{P}_{(X,Y)\sim\mathcal{D}_{XY}^{n_0}}\left(\mathrm{MutualNN}_{k,d_X,d_Y}(X,Y) = 0\right) \\
&= \mathbb{P}_{(X,Y)\sim\mathcal{D}_{XY}^{n_0}}\left(c(X)=0\right) \\
&= (1-\beta)^{n_0}
\end{aligned}
$$

□

E.4 P-value Bounds from Distribution-Free Tail Bounds

Theorem 2 (P-value Bounds from Distribution-Free Tail Bounds)**.** *Let S be a similarity metric with non-trivial distribution-free right-tail and left-tail bounds u_R and u_L, and consider the test of Definition 7 with margin $t \geq 0$. Under H_0, $S^*_{d_X,d_Y}(\mathcal{D}_{XY}) \leq S^*_{d_{X'},d_{Y'}}(\mathcal{D}_{X'Y'}) + t$, for an observed test statistic $T_S = \Delta > t$ and $N, N' \in \mathbb{N}$, the tail probability $\mathbb{P}_{H_0}(T_S \geq \Delta)$ satisfies:*

$$
\mathbb{P}_{H_0}(T_S \geq \Delta) \leq \inf_{0<\lambda<\Delta-t} \left[u_R(\lambda, N) + u_L(\Delta - t - \lambda, N')\right].
$$

Moreover, if the samples (X, Y) and (X', Y') are independent, then

$$\mathbb{P}_{H_0}(T_S \geq \Delta) \leq \inf_{0<\lambda<\Delta-t} \left[u_R(\lambda, N) + u_L(\Delta - t - \lambda, N') - u_R(\lambda, N)u_L(\Delta - t - \lambda, N')\right].$$

Proof. Recall from Definition 7 that the tail probability $\mathbb{P}_{H_0}(T_S \geq \Delta)$ is a shorthand for the joint sampling probability under H_0, which we abbreviate throughout the proof by

$$\mathbb{P} := \mathbb{P}_{\substack{(X,Y)\sim\mathcal{D}_{XY}^{N} \\ (X',Y')\sim\mathcal{D}_{X'Y'}^{N'}}}, \qquad \text{s. t.} \qquad \mathbb{P}_{H_0}(T_S \geq \Delta) = \mathbb{P}\big(S_{d_X,d_Y}(X, Y) - S_{d_{X'},d_{Y'}}(X', Y') \geq \Delta\big).$$

For brevity, write

$$\begin{aligned} A &= S_{d_X,d_Y}(X, Y), \\ B &= S_{d_{X'},d_{Y'}}(X', Y'), \\ A^* &= S^*_{d_X,d_Y}(\mathcal{D}_{XY}), \\ B^* &= S^*_{d_{X'},d_{Y'}}(\mathcal{D}_{X'Y'}). \end{aligned}$$

Let $r := \Delta + t + B^* - A^*$. By assumption, $r \geq \Delta$, and

$$\begin{aligned} &\mathbb{P}_{H_0}(T_S \geq \Delta) \\ &= \mathbb{P}(A - B \geq \Delta) \\ &= \mathbb{P}((A - A^*) + (B^* - B) \geq r - t). \end{aligned}$$

Fix any $\lambda \in (0, \Delta - t)$. If $A - A^* < \lambda$ and $B^* - B < r - t - \lambda$, then $(A - A^*) + (B^* - B) < r - t$. Hence, by union bound, the definitions of u_R and u_L and noting that $r - t \geq \Delta - t > \Delta - \Delta = 0$,

$$\begin{aligned} &\mathbb{P}((A - A^*) + (B^* - B) \geq r - t) \\ &\leq \mathbb{P}((A - A^* \geq \lambda) \vee (B^* - B \geq r - t - \lambda)) \\ &\leq \mathbb{P}(A - A^* \geq \lambda) + \mathbb{P}(B^* - B \geq r - t - \lambda) \\ &\leq u_R(\lambda, N) + u_L(r - t - \lambda, N'). \end{aligned}$$

Since $r - t - \lambda \geq \Delta - t - \lambda$ and u_L is non-increasing in its first argument, $u_L(r - t - \lambda, N') \leq u_L(\Delta - t - \lambda, N')$. Therefore,

$$\mathbb{P}(A - B \geq \Delta) \leq u_R(\lambda, N) + u_L(\Delta - t - \lambda, N'),$$

Taking the infimum over $\lambda \in (0, \Delta - t)$ gives the result.

Now consider the setting where (X, Y) and (X', Y') are independent, then $A - A^* \geq \lambda$ depends only on (X, Y) and $B^* - B \geq r - t - \lambda$ only on (X', Y'), so the two events are independent. Using the identity $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$, which under independence gives $\mathbb{P}(E \cap F) = \mathbb{P}(E)\mathbb{P}(F)$:

$$\begin{aligned} &\mathbb{P}((A - A^* \geq \lambda) \vee (B^* - B \geq r - t - \lambda)) \\ &= \mathbb{P}(A - A^* \geq \lambda) + \mathbb{P}(B^* - B \geq r - t - \lambda) - \mathbb{P}(A - A^* \geq \lambda)\,\mathbb{P}(B^* - B \geq r - t - \lambda) \\ &\leq u_R(\lambda, N) + u_L(r - t - \lambda, N') - u_R(\lambda, N)\,u_L(r - t - \lambda, N'). \end{aligned}$$

The last inequality holds because $p + q - pq = 1 - (1 - p)(1 - q)$ is non-decreasing in each of $p, q \in [0, 1]$, and u_R, u_L take values in $[0, 1]$ (Definition 6), so replacing each probability by its upper bound can only increase the right-hand side.

Since $r - t - \lambda \geq \Delta - t - \lambda$ and u_L is non-increasing in its first argument, $u_L(r - t - \lambda, N') \leq u_L(\Delta - t - \lambda, N')$. By the same monotonicity of $p + q - pq$ on $[0, 1]^2$, in the independent case

$$\mathbb{P}(A - B \geq \Delta) \leq u_R(\lambda, N) + u_L(\Delta - t - \lambda, N') - u_R(\lambda, N)\,u_L(\Delta - t - \lambda, N').$$

Taking the infimum over $\lambda \in (0, \Delta - t)$ gives the result. □

F Additional Proofs

F.1 Additional P-value Bounds from Distribution-Free Tail Bounds

Corollary 3 (Additional P-value Bounds from Distribution-Free Tail Bounds)**.** *Let S be a bounded similarity metric in $[0,1]$ with non-trivial distribution-free right-tail and left-tail bounds u_R and u_L, and consider the test of Definition 7 with margin $t \geq 0$. Under H_0, $S^*_{d_X,d_Y}(\mathcal{D}_{XY}) \leq S^*_{d_{X'},d_{Y'}}(\mathcal{D}_{X'Y'}) + t$ with (X,Y) and (X',Y') independent, for an observed test statistic $T_S = \Delta > t$ and $N, N' \in \mathbb{N}$, the tail probability $\mathbb{P}_{H_0}(T_S \geq \Delta)$ satisfies:*

$$\mathbb{P}_{H_0}(T_S \geq \Delta) \leq \inf_{\lambda \geq 0} \frac{\left(1 + \lambda \int_0^1 e^{\lambda s}\, u_R(s,N)\, ds\right)\left(1 + \lambda \int_0^1 e^{\lambda s}\, u_L(s,N')\, ds\right)}{e^{\lambda(\Delta - t)}}.$$

Proof. Recall from Definition 7 that $\mathbb{P}_{H_0}(T_S \geq \Delta)$ is a shorthand for the joint sampling probability under H_0, which we abbreviate throughout by

$$\mathbb{P} := \mathbb{P}_{\substack{(X,Y)\sim\mathcal{D}_{XY}^N \\ (X',Y')\sim\mathcal{D}_{X'Y'}^{N'}}}, \qquad \text{s. t.} \qquad \mathbb{P}_{H_0}(T_S \geq \Delta) = \mathbb{P}\big(S_{d_X,d_Y}(X,Y) - S_{d_{X'},d_{Y'}}(X',Y') \geq \Delta\big).$$

For brevity, write

$$\begin{aligned} A &= S_{d_X,d_Y}(X,Y), \\ B &= S_{d_{X'},d_{Y'}}(X',Y'), \\ A^* &= S^*_{d_X,d_Y}(\mathcal{D}_{XY}), \\ B^* &= S^*_{d_{X'},d_{Y'}}(\mathcal{D}_{X'Y'}). \end{aligned}$$

Let $r := \Delta + t + B^* - A^*$. By assumption, $r \geq \Delta$, and

$$\begin{aligned} &\mathbb{P}_{H_0}(T_S \geq \Delta) \\ &= \mathbb{P}(A - B \geq \Delta) \\ &= \mathbb{P}((A - A^*) + (B^* - B) \geq r - t). \end{aligned}$$

Furthermore, for any $\lambda \geq 0$, using additional linear algebra tricks and Markov's Inequality we get:

$$\begin{aligned} &\mathbb{P}((A - A^*) + (B^* - B) \geq r - t) \\ &\leq \mathbb{P}((A - A^*) + (B^* - B) \geq \Delta - t) \\ &= \mathbb{P}\Big(e^{\lambda(A-A^*)+\lambda(B^*-B)} \geq e^{\lambda(\Delta-t)}\Big) \\ &\leq \frac{\mathbb{E}[e^{\lambda(A-A^*)+\lambda(B^*-B)}]}{e^{\lambda(\Delta-t)}} \\ &= \frac{\mathbb{E}[e^{\lambda(A-A^*)}e^{\lambda(B^*-B)}]}{e^{\lambda(\Delta-t)}} \\ &= \frac{\mathbb{E}[e^{\lambda(A-A^*)}]\mathbb{E}[e^{\lambda(B^*-B)}]}{e^{\lambda(\Delta-t)}} \end{aligned}$$

It remains to bound $\mathbb{E}[e^{\lambda(A-A^*)}]$ and $\mathbb{E}[e^{\lambda(B^*-B)}]$ by the tail bounds u_R and u_L. We assume the similarity scores take values in $[0,1]$. Recall that for any nonnegative random variable W, the following identity holds for the moment generating function:

$$\mathbb{E}[e^{\lambda W}] = 1 + \lambda \int_0^{+\infty} e^{\lambda s}\, \mathbb{P}(W \geq s)\, ds.$$

Write $W \coloneqq (A - A^*)_+$. Then $W \in [0,1]$ and $e^{\lambda(A-A^*)} \le e^{\lambda W}$, so $\mathbb{E}[e^{\lambda(A-A^*)}] \le \mathbb{E}[e^{\lambda W}]$. As a consequence, we get that for $\varepsilon > 0$:

$$\begin{aligned}
&\mathbb{E}[e^{\lambda(A-A^*)}] \\
&\le \mathbb{E}[e^{\lambda W}] \\
&= 1 + \lambda \int_0^{+\infty} e^{\lambda s}\, \mathbb{P}(W \ge s)\, ds \\
&= 1 + \lambda \int_0^{1} e^{\lambda s}\, \mathbb{P}(W \ge s)\, ds \\
&= 1 + \lambda \int_\varepsilon^{1} e^{\lambda s}\, \mathbb{P}(W \ge s)\, ds + \lambda \int_0^{\varepsilon} e^{\lambda s}\, \mathbb{P}(W \ge s)\, ds \\
&\le 1 + \lambda \int_\varepsilon^{1} e^{\lambda s}\, u_R(s,N)\, ds + \lambda \int_0^{\varepsilon} e^{\lambda s}\, \mathbb{P}(W \ge s)\, ds
\end{aligned}$$

The remainder satisfies $\lambda \int_0^\varepsilon e^{\lambda s}\, \mathbb{P}(W \ge s)\, ds \le \lambda \varepsilon e^{\lambda \varepsilon} \to 0$ as $\varepsilon \to 0$, and therefore

$$\mathbb{E}[e^{\lambda(A-A^*)}] \le 1 + \lambda \int_0^1 e^{\lambda s}\, u_R(s,N)\, ds$$

The same argument applies to B. Write $Z \coloneqq (B^* - B)_+$. Then $Z \in [0,1]$ and $e^{\lambda(B^*-B)} \le e^{\lambda Z}$. For every $s \in (0,1]$, $\mathbb{P}(Z \ge s) = \mathbb{P}(B^* - B \ge s) \le u_L(s,N')$, and therefore

$$\mathbb{E}[e^{\lambda(B^*-B)}] \le 1 + \lambda \int_0^1 e^{\lambda s}\, u_L(s,N')\, ds.$$

Substituting these two bounds into the Markov estimate and taking the infimum over $\lambda \ge 0$ yields

$$\mathbb{P}(A - B \ge \Delta) \le \inf_{\lambda \ge 0} \frac{\big(1 + \lambda \int_0^1 e^{\lambda s}\, u_R(s,N)\, ds\big)\big(1 + \lambda \int_0^1 e^{\lambda s}\, u_L(s,N')\, ds\big)}{e^{\lambda(\Delta - t)}}$$

□

F.2 Closed-Form P-value Bounds under Sub-Gaussian Tail Bounds

Corollary 4 (Closed-Form P-value Bounds under Sub-Gaussian Tail Bounds)**.** *Let S be a bounded similarity metric in $[0,1]$ with distribution-free tail bounds of the form*

$$u_R(\varepsilon, N) = u_L(\varepsilon, N) = a \exp\big(-g(N)\varepsilon^2\big),$$

where $a > 0$ and $g : \mathbb{N} \to \mathbb{R}_+$ satisfies $g(N) > 0$. Consider the test of Definition 7 with margin $t \ge 0$. Under H_0, with (X,Y) and (X',Y') independent, an observed test statistic $T_S = \Delta > t$, and $N, N' \in \mathbb{N}$,

$$\mathbb{P}_{H_0}(T_S \ge \Delta) \le \Big(1 + a\sqrt{\pi}\, e^{1/4}\Big)^2 \exp\Big(-\min\big\{\sqrt{g(N)}, \sqrt{g(N')}\big\}(\Delta - t)\Big).$$

Proof. By Corollary 3, we have that under H_0:

$$\begin{aligned}
&\mathbb{P}_{\substack{(X,Y)\sim\mathcal{D}_{XY}^N \\ (X',Y')\sim\mathcal{D}_{X'Y'}^{N'}}} \big(S_{d_X,d_Y}(X,Y) - S_{d_{X'},d_{Y'}}(X',Y') \ge \Delta\big) \\
&\le \inf_{\lambda \ge 0} \frac{\big(1 + \lambda \int_0^1 e^{\lambda s}\, u_R(s,N)\, ds\big)\big(1 + \lambda \int_0^1 e^{\lambda s}\, u_L(s,N')\, ds\big)}{e^{\lambda(\Delta - t)}}
\end{aligned}$$

Since $T_S = S_{d_X,d_Y}(X,Y) - S_{d_{X'},d_{Y'}}(X',Y')$, it follows under H_0 that

$$\begin{aligned}
&\mathbb{P}_{H_0}(T_S \ge \Delta) \\
&= \mathbb{P}_{\substack{(X,Y)\sim\mathcal{D}_{XY}^N \\ (X',Y')\sim\mathcal{D}_{X'Y'}^{N'}}} \big(S_{d_X,d_Y}(X,Y) - S_{d_{X'},d_{Y'}}(X',Y') \ge \Delta\big) \\
&\le \inf_{\lambda \ge 0} \frac{\big(1 + \lambda \int_0^1 e^{\lambda s}\, u_R(s,N)\, ds\big)\big(1 + \lambda \int_0^1 e^{\lambda s}\, u_L(s,N')\, ds\big)}{e^{\lambda(\Delta - t)}}.
\end{aligned}$$

Assume that a given similarity metric admits tail bounds of the form

$$u_R(\varepsilon, N) = u_L(\varepsilon, N) = a \cdot \exp(-g(N)\varepsilon^2)$$

Write the objective inside the infimum as $F(\lambda)$ by the definition of the infimum,

$$\mathbb{P}_{H_0}(T_S \geq \Delta) \leq \inf_{\lambda \geq 0} F(\lambda) \leq F(\min(\sqrt{g(N)}, \sqrt{g(N')})).$$

Thus, evaluating the objective at $\lambda = \min(\sqrt{g(N)}, \sqrt{g(N')})$ gives a valid explicit upper bound.

We first simplify the objective:

$$\begin{aligned}
&\inf_{\lambda \geq 0} \frac{\left(1 + \lambda \int_0^1 e^{\lambda s}\, u_R(s, N)\, ds\right)\left(1 + \lambda \int_0^1 e^{\lambda s}\, u_L(s, N')\, ds\right)}{e^{\lambda(\Delta - t)}} \\
&= \inf_{\lambda \geq 0} \frac{\left(1 + \lambda \int_0^1 e^{\lambda s}\, a \cdot e^{-g(N)s^2}\, ds\right)\left(1 + \lambda \int_0^1 e^{\lambda s}\, a \cdot e^{-g(N')s^2}\, ds\right)}{e^{\lambda(\Delta - t)}} \\
&= \inf_{\lambda \geq 0} \frac{\left(1 + \lambda a \int_0^1 e^{\lambda s}\, e^{-g(N)s^2}\, ds\right)\left(1 + \lambda a \int_0^1 e^{\lambda s}\, e^{-g(N')s^2}\, ds\right)}{e^{\lambda(\Delta - t)}} \\
&= \inf_{\lambda \geq 0} \frac{\left(1 + \lambda a \int_0^1 e^{\lambda s - g(N)s^2}\, ds\right)\left(1 + \lambda a \int_0^1 e^{\lambda s - g(N')s^2}\, ds\right)}{e^{\lambda(\Delta - t)}}
\end{aligned}$$

Now we compute $\int_0^1 e^{\lambda s - g(N)s^2}\, ds$:

$$\begin{aligned}
&\int_0^1 e^{\lambda s - g(N)s^2}\, ds \\
&= \int_0^1 e^{\frac{\lambda^2}{4g(N)} - \left(\sqrt{g(N)}s - \frac{\lambda}{2\sqrt{g(N)}}\right)^2}\, ds
\end{aligned}$$

Substitute $u = \frac{2g(N)s - \lambda}{2\sqrt{g(N)}} \implies du = \sqrt{g(N)}\, ds$, with bounds $u(0) = -\frac{\lambda}{2\sqrt{g(N)}}$ and $u(1) = \frac{2g(N) - \lambda}{2\sqrt{g(N)}}$:

$$= \frac{\sqrt{\pi} e^{\frac{\lambda^2}{4g(N)}}}{2\sqrt{g(N)}} \int_{-\frac{\lambda}{2\sqrt{g(N)}}}^{\frac{2g(N)-\lambda}{2\sqrt{g(N)}}} \frac{2}{\sqrt{\pi}} e^{-u^2}\, du$$

Using the Gauss error function $\int \frac{2}{\sqrt{\pi}} e^{-u^2}\, du = \operatorname{erf}(u)$:

$$\begin{aligned}
&= \frac{\sqrt{\pi} e^{\frac{\lambda^2}{4g(N)}}}{2\sqrt{g(N)}} \left[\operatorname{erf}(u)\right]_{-\frac{\lambda}{2\sqrt{g(N)}}}^{\frac{2g(N)-\lambda}{2\sqrt{g(N)}}} \\
&= \frac{\sqrt{\pi} e^{\frac{\lambda^2}{4g(N)}}}{2\sqrt{g(N)}} \left(\operatorname{erf}\left(\frac{2g(N) - \lambda}{2\sqrt{g(N)}}\right) - \operatorname{erf}\left(-\frac{\lambda}{2\sqrt{g(N)}}\right)\right) \\
&= \frac{\sqrt{\pi} e^{\frac{\lambda^2}{4g(N)}}}{2\sqrt{g(N)}} \left(\operatorname{erf}\left(\frac{2g(N) - \lambda}{2\sqrt{g(N)}}\right) + \operatorname{erf}\left(\frac{\lambda}{2\sqrt{g(N)}}\right)\right).
\end{aligned}$$

Given that for all $x \in \mathbb{R} : \operatorname{erf}(x) \leq 1$, we see that:

$$\int_0^1 e^{\lambda s - g(N)s^2}\, ds \leq \frac{\sqrt{\pi} e^{\frac{\lambda^2}{4g(N)}}}{\sqrt{g(N)}}$$

Replacing this into the original equation, we see that:

$$\begin{aligned}
&\inf_{\lambda \geq 0} \frac{\left(1 + \lambda a \int_0^1 e^{\lambda s - g(N)s^2}\, ds\right)\left(1 + \lambda a \int_0^1 e^{\lambda s - g(N')s^2}\, ds\right)}{e^{\lambda(\Delta - t)}} \\
&\leq \inf_{\lambda \geq 0} \frac{\left(1 + \lambda a \frac{\sqrt{\pi} e^{\frac{\lambda^2}{4g(N)}}}{\sqrt{g(N)}}\right)\left(1 + \lambda a \frac{\sqrt{\pi} e^{\frac{\lambda^2}{4g(N')}}}{\sqrt{g(N')}}\right)}{e^{\lambda(\Delta - t)}}
\end{aligned}$$

Evaluating at $\lambda = \min\{\sqrt{g(N)}, \sqrt{g(N')}\}$ gives:

$$\frac{\left(1 + a\sqrt{\pi}\min(\sqrt{g(N)}, \sqrt{g(N')})\frac{e^{\frac{\min(g(N),g(N'))}{4g(N)}}}{\sqrt{g(N)}}\right)\left(1 + a\sqrt{\pi}\min(\sqrt{g(N)}, \sqrt{g(N')})\frac{e^{\frac{\min(g(N),g(N'))}{4g(N')}}}{\sqrt{g(N')}}\right)}{e^{\min(\sqrt{g(N)},\sqrt{g(N')})(\Delta - t)}}$$

$$\leq \frac{\left(1 + a\sqrt{\pi}\min(\sqrt{g(N)}, \sqrt{g(N')})\frac{e^{\frac{1}{4}}}{\sqrt{g(N)}}\right)\left(1 + a\sqrt{\pi}\min(\sqrt{g(N)}, \sqrt{g(N')})\frac{e^{\frac{1}{4}}}{\sqrt{g(N')}}\right)}{e^{\min(\sqrt{g(N)},\sqrt{g(N')})(\Delta - t)}}$$

$$\leq \frac{\left(1 + \frac{a\sqrt{\pi}\min(\sqrt{g(N)},\sqrt{g(N')})\ e^{\frac{1}{4}}}{\min(\sqrt{g(N)},\sqrt{g(N')})}\right)\left(1 + \frac{a\sqrt{\pi}\min(\sqrt{g(N)},\sqrt{g(N')})\ e^{\frac{1}{4}}}{\min(\sqrt{g(N)},\sqrt{g(N')})}\right)}{e^{\min(\sqrt{g(N)},\sqrt{g(N')})(\Delta - t)}}$$

$$= \frac{\left(1 + a\sqrt{\pi}e^{\frac{1}{4}}\right)^2}{e^{\min(\sqrt{g(N)},\sqrt{g(N')})(\Delta - t)}}$$

$$= \left(1 + a\sqrt{\pi}e^{\frac{1}{4}}\right)^2 e^{-\min(\sqrt{g(N)},\sqrt{g(N')})(\Delta - t)}$$

□

F.3 Boundedness of CKA-RBF

Lemma 1 (Boundedness of CKA-RBF)**.** *Under Assumption 2, empirical* CKA-RBF *satisfies*

$$\text{CKA-RBF}_{\sigma_X,\sigma_Y}(X, Y) \in [0, 1].$$

Proof. By Definition 2, empirical CKA-RBF is defined as the normalized HSIC. Using the Frobenius inner product $\langle A, B\rangle_F = \operatorname{tr}(AB)$ for symmetric matrices and its induced norm $\|A\|_F = \sqrt{\operatorname{tr}(A^2)}$, we write:

$$\text{CKA-RBF}_{\sigma_X,\sigma_Y}(X, Y) = \frac{\text{HSIC}_{\sigma_X,\sigma_Y}(X, Y)}{\sqrt{\text{HSIC}_{\sigma_X,\sigma_X}(X, X)\,\text{HSIC}_{\sigma_Y,\sigma_Y}(Y, Y)}} = \frac{\langle K_X^c, K_Y^c\rangle_F}{\|K_X^c\|_F\|K_Y^c\|_F},$$

where $K_X^c = HK_XH$ and $K_Y^c = HK_YH$ are the centered Gram matrices with centering matrix $H = I_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$. By Theorem 9, the CND condition ensures that K_X^c and K_Y^c are symmetric and positive semi-definite (PSD):

(i) **Upper Bound** (≤ 1)**:** By the Cauchy-Schwarz inequality for the Frobenius inner product (Theorem 8),

$$\langle K_X^c, K_Y^c\rangle_F \leq \|K_X^c\|_F\|K_Y^c\|_F \implies \text{CKA-RBF}_{\sigma_X,\sigma_Y}(X, Y) \leq 1.$$

(ii) **Lower Bound** (≥ 0)**:** Because K_X^c and K_Y^c are symmetric and PSD, K_X^c admits a unique symmetric square root $(K_X^c)^{1/2}$. By the cyclic property of the trace:

$$\operatorname{tr}(K_X^cK_Y^c) = \operatorname{tr}\Big((K_X^c)^{1/2}K_Y^c(K_X^c)^{1/2}\Big) \geq 0,$$

since $(K_X^c)^{1/2}K_Y^c(K_X^c)^{1/2}$ is also PSD and therefore has non-negative trace. Hence, $\text{CKA-RBF}_{\sigma_X,\sigma_Y}(X, Y) \geq 0$. □

F.4 Boundedness of HSIC

Lemma 2 (Boundedness of HSIC)**.** *Let $k_X : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $k_Y : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be bounded kernels satisfying $|k_X(x, x')| \leq c_x$ and $|k_Y(y, y')| \leq c_y$ for all $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$. Then, the empirical estimator* $\text{HSIC}_{k_X,k_Y}(X, Y)$ *is bounded.*

Proof. Following Gretton et al. (2007), the empirical estimator $\mathrm{HSIC}_{k_X,k_Y}(X,Y)$ can be expressed as a V-statistic of degree $m=4$ over the paired observations $z_i=(x_i,y_i)\sim\mathcal{D}_{XY}$ with symmetric kernel:

$$h(z_i,z_j,z_q,z_r)=\frac{1}{4!}\sum_{(t,u,v,w)}^{(i,j,q,r)}\Big(k_{tu}l_{tu}+k_{tu}l_{vw}-2k_{tu}l_{tv}\Big),$$

where $k_{ij}=k_X(x_i,x_j)$, $l_{ij}=k_Y(y_i,y_j)$, and the summation indexes all ordered quadruples (t,u,v,w) drawn without replacement from the index set (i,j,q,r).

Applying the triangle inequality directly to the kernel expression yields:

$$\begin{aligned}|h(z_i,z_j,z_q,z_r)|&\le\frac{1}{4!}\sum_{(t,u,v,w)}^{(i,j,q,r)}\Big(|k_{tu}l_{tu}|+|k_{tu}l_{vw}|+2|k_{tu}l_{tv}|\Big)\\&\le\frac{1}{4!}\sum_{(t,u,v,w)}^{(i,j,q,r)}\Big(c_xc_y+c_xc_y+2c_xc_y\Big)\\&=\frac{1}{24}\cdot24\cdot4c_xc_y=4c_xc_y.\end{aligned}$$

Because the empirical estimator $\mathrm{HSIC}_{k_X,k_Y}(X,Y)$ is defined as an average of evaluations of the kernel h, applying the triangle inequality across all sample quadruples immediately gives:

$$|\mathrm{HSIC}_{k_X,k_Y}(X,Y)|\le\frac{1}{N^4}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{q=1}^{N}\sum_{r=1}^{N}|h(z_i,z_j,z_q,z_r)|\le4c_xc_y,$$

which completes the proof. □

F.5 Lower Bound Preservation for Empirical HSIC

Lemma 3 (Lower Bound Preservation for Empirical HSIC)**.** *Let k_X be a positive semi-definite kernel, and let $m\in\mathbb{R}$. Suppose that for every $N\ge2$,*

$$\mathbb{P}\big(\mathrm{HSIC}_{k_X,k_X}(X,X)\ge m\big)=1.$$

If $\mathrm{HSIC}_{k_X,k_X}(X,X)$ converges in probability to the deterministic constant $\mathrm{HSIC}^*_{k_X,k_X}(\mathcal{D}_{XX})$ as $N\to\infty$, then:

$$\mathrm{HSIC}^*_{k_X,k_X}(\mathcal{D}_{XX})\ge m.$$

Proof. We proceed by contradiction. Assume that $\mathrm{HSIC}^*_{k_X,k_X}(\mathcal{D}_{XX})<m$, and define $\varepsilon_0=m-\mathrm{HSIC}^*_{k_X,k_X}(\mathcal{D}_{XX})>0$.

By the definition of convergence in probability, for any $\varepsilon>0$,

$$\lim_{N\to\infty}\mathbb{P}\Big(\big|\mathrm{HSIC}_{k_X,k_X}(X,X)-\mathrm{HSIC}^*_{k_X,k_X}(\mathcal{D}_{XX})\big|\ge\varepsilon\Big)=0.$$

Evaluating this convergence at $\varepsilon=\varepsilon_0$ gives:

$$\lim_{N\to\infty}\mathbb{P}\Big(\big|\mathrm{HSIC}_{k_X,k_X}(X,X)-\mathrm{HSIC}^*_{k_X,k_X}(\mathcal{D}_{XX})\big|\ge\varepsilon_0\Big)=0.$$

Whenever $\mathrm{HSIC}_{k_X,k_X}(X,X)\ge m$, we have:

$$\mathrm{HSIC}_{k_X,k_X}(X,X)-\mathrm{HSIC}^*_{k_X,k_X}(\mathcal{D}_{XX})\ge m-\mathrm{HSIC}^*_{k_X,k_X}(\mathcal{D}_{XX})=\varepsilon_0>0,$$

which implies $\big|\mathrm{HSIC}_{k_X,k_X}(X,X)-\mathrm{HSIC}^*_{k_X,k_X}(\mathcal{D}_{XX})\big|\ge\varepsilon_0$. Consequently, the event $\{\mathrm{HSIC}_{k_X,k_X}(X,X)\ge m\}$ is a subset of the deviation event:

$$\big\{\mathrm{HSIC}_{k_X,k_X}(X,X)\ge m\big\}\subseteq\Big\{\big|\mathrm{HSIC}_{k_X,k_X}(X,X)-\mathrm{HSIC}^*_{k_X,k_X}(\mathcal{D}_{XX})\big|\ge\varepsilon_0\Big\}.$$

By the monotonicity of probability measures and our almost sure assumption:

$$1 = \mathbb{P}\big(\mathrm{HSIC}_{k_X,k_X}(X,X) \geq m\big) \leq \mathbb{P}\Big(\big|\mathrm{HSIC}_{k_X,k_X}(X,X) - \mathrm{HSIC}^*_{k_X,k_X}(\mathcal{D}_{XX})\big| \geq \varepsilon_0\Big),$$

for all $N \geq 2$. Taking the limit as $N \to \infty$ yields:

$$1 \leq \lim_{N\to\infty} \mathbb{P}\Big(\big|\mathrm{HSIC}_{k_X,k_X}(X,X) - \mathrm{HSIC}^*_{k_X,k_X}(\mathcal{D}_{XX})\big| \geq \varepsilon_0\Big) = 0,$$

which is a contradiction. Therefore, we must have $\mathrm{HSIC}^*_{k_X,k_X}(\mathcal{D}_{XX}) \geq m$. □