A UNIFIED GEOMETRIC FRAMEWORK FOR INFORMATION-THEORETIC ANALYSIS AND DECOMPOSITION

RESEARCH PROPOSAL

Zhaozhao Ma[⋆,†]
[⋆]*Zhejiang University*
[†]*Georgia Institute of Technology*
zhaozhaoma@{zju.edu.cn, gatech.edu}

ABSTRACT

Given a frozen text encoder f that produces high-dimensional embeddings $\mathbf{Z} = f(X) \in \mathbb{R}^d$, direct Mutual Information (MI) estimation suffers from loose bounds, instability, and incomparability across different models. We propose a unified framework for information-theoretic analysis based on Centered Kernel Alignment (CKA), which provides a robust, geometrically-grounded metric for measuring dependencies in high-dimensional representation spaces.

1 Introduction

Each new generation of embedding models is accompanied by refreshed leaderboard results on benchmarks such as MTEB, yet these evaluations remain partial, non-necessary, and sometimes misleading (see Sec. 2). A central practical question thus arises: how can we reliably assess embedding quality, especially for domain-specific tasks? For example, general benchmarks like MTEB contain almost no financial tasks; consequently, a model's ranking on MTEB does not necessarily translate to usefulness for financial text encoding. Similar limitations appear in law and biomedicine, where domain semantics and terminology dominate.

These challenges motivate an evaluation paradigm that avoids predefined benchmarks and instead directly measures an embedding's effectiveness for a given dataset and prediction target. Accordingly, this proposal introduces an information-theoretic evaluation framework that quantifies how much an embedding captures from two fundamental sources: the raw input X and the target Y. Using logarithmic CKA, we define information measures L_X and L_Y (dependence with X or Y) and residuals, R_X and R_Y (potentially missing information). This naturally yields a bottleneck-learning objective: with the embedding model frozen, we learn a low-dimensional projection that maximizes L_X and L_Y while minimizing R_X and R_Y. The resulting bottleneck representation isolates the most task-informative components of the embedding and provides a principled basis for comparing models.

Beyond single-model evaluation, the framework also supports two-source information decomposition, separating shared from unique information across embeddings. This allows practitioners to determine whether a new embedding model offers genuinely complementary information or merely overlaps with an existing one, directly informing fusion, ensembling, and distillation strategies.

In summary, this proposal aims to deliver:

- a unified information-theoretic metric of embedding quality;
- a training framework for learning optimal low-dimensional bottleneck representations under this metric;
- a decomposition perspective for comparing embeddings by quantifying their shared and unique information.

2 Literature Review

Evaluating Learned Embeddings. Embedding evaluation is typically grouped into: downstream task performance, benchmark-based holistic assessments, and intrinsic methods. Early sentence and vision embeddings (e.g., InferSent) were mainly evaluated via downstream tasks, under the assumption that stronger downstream performance yield better representations. However, such evaluations are inherently task-dependent, lack interpretability, and cannot disentangle

the intrinsic quality of the embeddings rendering them necessary but insufficient. Subsequent work introduced large scale benchmarks such as MTEB and BEIR to approximate true generalization. Yet these benchmarks suffer from high computational cost, limited interpretability, and increasing risks of data contamination and benchmark saturation. Notably, [] show that the top nine systems on French MTEB are statistically indistinguishable, introducing further doubt about whether benchmark rankings reflect meaningful performance gaps. Building on such recognition that fragmented task-level performance does not equate to true representation quality, researchers have explored more direct measures grounded in geometric structure and statistical dependence. For example, [**?**] examine local and global geometric properties of embeddings through manifold and neighborhood structure, while [**?**] analyze intrinsic dimensionality to detect potential information compression or redundancy within the embedding manifold.

Information Theoretic Measures for Embeddings. Information-theoretic approaches provide a rigorous and unified foundation for understanding and evaluating representation learning. Information Bottleneck formalize the classical compression–prediction trade-off by maximizing task-relevant information I(Y,Z) while minimizing input redundancy I(X,Z), where Y is the target, Z the embedding, and X the raw input, its subsequent extensions alleviate the difficulty of mutual-information estimation through variational approximations [**?**], enabling scalable, trainable neural implementations that have rapidly generalized across domains. HSIC provides a density-free alternative to mutual information by quantifying statistical dependence via the Hilbert–Schmidt norm of the RKHS cross-covariance operator. Its normalized variant, CKA, introduces invariance to isotropic scaling and layer dimensionality, enabling reliable comparison of representation spaces. Such dependence-alignment measures have been broadly utilized, from diagnosing redundancy in deep networks to comparing learned sentence representations and facilitating knowledge distillation.

Significance of Cross-Model Information Perspective. For model compression, transfer learning, and interpretability, quantifying how representations overlap or complement each other is of central importance. Methods such as Deep CCA [**??**] learn shared latent factors by maximizing cross-view correlation, thereby characterizing common structure across modalities. However, alignment-based approaches emphasize the shared component and cannot disentangle shared information from view-specific, task-relevant structure. Concurrent findings in deep representation learning further show that models trained on the same data may capture distinct feature subspaces—differing in abstraction level, frequency, or semantic direction [].

This raises structural questions: To what extent do two models capture overlapping information about a target Y? Does either model possess unique information unavailable to the other? Evidence from model fusion and knowledge distillation indicates that complementary task-relevant structures can enhance performance [**???**], whereas unnecessary redundancy can reduce efficiency or induce overfitting [**?**].

Information theory addresses these questions through the Partial Information Decomposition (PID) framework [**???**], which decomposes the contributions of two sources into shared, unique, and synergistic components. Yet PID remains notoriously difficult to estimate—particularly for high-dimensional continuous representations—and a scalable, interpretable, and information-theoretically consistent decomposition method is still lacking.

3 Methodology

3.1 Data Protocol

Given a frozen embedding model f that produces high dimensional embeddings $\mathbf{Z} = f(X) \in \mathbb{R}^a$ (where a can be 384, 768, or more). To ensure fair comparison across different f, define:

$$\mathbf{U} = T_x(X) \in \mathbb{R}^b, \tag{1}$$

where T_x is non-trainable transformation (e.g., TF-IDF followed by TruncatedSVD) maps raw input X (e.g., text data) into a fixed, low-dimensional representation $\mathbf{U}^b$ to ensure consistent input features and prevent a learnable encoder from absorbing or masking specific model differences.

3.2 The Kernel Alignment Dependence Metric

3.2.1 Defining Dependence via HSIC and CKA

Definition 1 (HSIC [Gretton et al., 2005])**.** HSIC measures dependence based on Hilbert-Schmidt norm of the centered cross-covariance operator between two variables, U and V, in their respective RKHSs. Given centered Gram matrices $\tilde{\mathbf{K}}$ and $\tilde{\mathbf{L}}$ (where $\tilde{\mathbf{K}} = \mathbf{HKH}$, $\tilde{\mathbf{L}} = \mathbf{HLH}$ and $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\mathsf{T}$), the empirical HSIC estimator is:

$$\mathrm{HSIC}(\mathbf{K}, \mathbf{L}) = \frac{1}{(n-1)^2}\,\mathrm{tr}(\tilde{\mathbf{K}}\tilde{\mathbf{L}}), \tag{2}$$

for simplicity, use $\langle \tilde{\mathbf{K}}, \tilde{\mathbf{L}} \rangle_F$, $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product, as the unnormalized statistic, which is equivalent up to scaling factors.

Definition 2 (CKA Kornblith et al. [2019])**.** CKA normalizes HSIC to produce a cosine similarity metric between kernel matrices, making it invariant to scaling:

$$\mathrm{CKA}(\mathbf{K}, \mathbf{L}) = \frac{\langle \tilde{\mathbf{K}}, \tilde{\mathbf{L}} \rangle_F}{\sqrt{\langle \tilde{\mathbf{K}}, \tilde{\mathbf{K}} \rangle_F \langle \tilde{\mathbf{L}}, \tilde{\mathbf{L}} \rangle_F}}. \tag{3}$$

$\mathrm{CKA} \in [0, 1]$, with 0 indicating independence and 1 fully aligned.

3.2.2 Unified Log-CKA Metric

Definition 3 (Unified Dependence Metric)**.** Define a unified information-theoretic metric $\hat{I}_{\mathrm{dep}}$ as a monotonic logarithmic transform of Eq. 3, which converts the measure from $[0, 1]$-bounded similarity score into unbounded dependence measure analogous to mutual information:

$$\widehat{I}_{\mathrm{dep}}(\mathbf{K}, \mathbf{L}) = -\log \frac{\langle \tilde{\mathbf{K}}, \tilde{\mathbf{L}} \rangle_F^2}{\langle \tilde{\mathbf{K}}, \tilde{\mathbf{K}} \rangle_F \langle \tilde{\mathbf{L}}, \tilde{\mathbf{L}} \rangle_F} = -2 \log \left(\mathrm{CKA}(\mathbf{K}, \mathbf{L}) \right). \tag{4}$$

Remark 1. It is critical to distinguish Eq. 4 from classical Parzen-window estimator form Jenssen et al. [2006], Principe [2010], such estimators are based on non-centered kernel sums (approximating density integrals like $\int p^2, \int pq$) and highly sensitive to density, scale, or shift. Eq. 4, by definition, based on centered kernels and CKA, which makes it invariant to translation and isotropic scaling, rendering it far more robust for the task of comparing neural representations Kornblith et al. [2019]. Detailed discussions of statistical properties including consistency, finite sample bias, and asymptotic normality, are provided in Appendix , Lemma 2, 3, and 4, respectively.

3.3 Bridging Dimensional Gaps via Orthogonal Residuals

3.3.1 Lower Bounds: Low-Dimensional Principal Terms

Define principal dependence terms captured in low-dimensional bottleneck space $\mathbf{V}$ (with Gram matrix $\mathbf{K}_V$) using Eq. 4:

$$L_X = \widehat{I}_{\mathrm{dep}}(\mathbf{K}_X, \mathbf{K}_V), \quad L_Y = \widehat{I}_{\mathrm{dep}}(\mathbf{K}_Y, \mathbf{K}_V). \tag{5}$$

3.3.2 Geometric Residuals

To estimate the information not captured by $\mathbf{V}$, this proposal introduces a geometric residual method. Instead of estimating complex conditional densities (e.g., $p(X|\mathbf{V})$), the method geometrically project the high-dimensional kernel $\mathbf{K}_Z$ onto the subspace orthogonal to that spanned by $\mathbf{V}$. For computational efficiency, Random Fourier Features (RFF) Rahimi and Recht [2007] are employed. Let $\Phi(\mathbf{V}) \in \mathbb{R}^{n \times m}$ denote the RFF matrix of $\mathbf{V}$, QR decomposition is then performed:

$$\Phi = \mathbf{QR}, \quad \text{where} \quad \mathbf{Q}^\top \mathbf{Q} = \mathbf{I}. \tag{6}$$

The projection onto the subspace spanned by $\mathbf{V}$ is given by $\Pi = \mathbf{QQ}^\top$, while the projection onto its orthogonal complement is $\mathbf{M} = \mathbf{I} - \Pi$. A residual Gram matrix $\mathbf{G}_{\mathrm{res}}$ is constructed by projecting $\mathbf{K}_Z$ onto the orthogonal subspace:

$$\mathbf{G}_{\mathrm{res}} = \mathbf{M}\mathbf{K}_Z\mathbf{M}. \tag{7}$$

The residual dependence, serving as a surrogate for conditional dependence, is then defined by applying the same unified metric Eq. 4 to the centered residual kernel:

$$R_X = \widehat{I}_{\mathrm{dep}}(\mathbf{K}_X, \tilde{\mathbf{G}}_{\mathrm{res}}), \quad R_Y = \widehat{I}_{\mathrm{dep}}(\mathbf{K}_Y, \tilde{\mathbf{G}}_{\mathrm{res}}). \tag{8}$$

where $\tilde{\mathbf{G}}_{\mathrm{res}} = \mathbf{H}\mathbf{G}_{\mathrm{res}}\mathbf{H}$ denotes the centered form of the residual kernel. The convergence and error analysis of geometric residuals can be found in Appendix A.2.

3.3.3 Upper Bounds: Log-Domain Conservative Term

Geometric definitions of L_X and R_X lead to an elegant conservative upper bound U for the total information, let $L_X = \widehat{I}_{\mathrm{dep}}(\mathbf{K}_X, \Pi\mathbf{K}_Z\Pi)$ and $R_X = \widehat{I}_{\mathrm{dep}}(\mathbf{K}_X, \mathbf{M}\mathbf{K}_Z\mathbf{M})$. Define:

$$U = \log(e^{L_X} + e^{R_X}), \tag{9}$$

$$\mathrm{gap} = U - L_X = \log(1 + e^{R_X - L_X}). \tag{10}$$

A small gap ($R_X \ll L_X$) implies $\mathbf{V}$ retains most of $\mathbf{Z}$'s geometric dependence, whereas a large gap ($R_X \gg L_X$) indicates that the bottleneck $\mathbf{V}$ fails to capture it.

Theorem 1 (Log-sum-exp upper bound)**.** *Let A be any PSD matrix (e.g., $A = \mathbf{K}_X$), and set $B_\parallel = \Pi\mathbf{K}_Z\Pi$, $B_\perp = M\mathbf{K}_Z M$, $B = B_\parallel + B_\perp$. Then:*

$$\widehat{I}_{\mathrm{dep}}(A, B) \;\le\; \log\Big(e^{\widehat{I}_{\mathrm{dep}}(A, B_\parallel)} + e^{\widehat{I}_{\mathrm{dep}}(A, B_\perp)} \Big). \tag{11}$$

Consequently, with $L_X = \widehat{I}_{\mathrm{dep}}(\mathbf{K}_X, B_\parallel)$ and $R_X = \widehat{I}_{\mathrm{dep}}(\mathbf{K}_X, B_\perp)$, we have $\widehat{I}_{\mathrm{dep}}(\mathbf{K}_X, B) \le U$. For completeness, please refer to Appendix A.3 for additional proof.

3.4 Training and Evaluation

The training objective is to find a optimal projection $T(\mathbf{Z}) = \mathbf{V}$ that maximizes the principal dependence terms (L_X, L_Y) while minimizing the residuals (R_X, R_Y). Formally, the overall training loss of the framework can be written as:

$$\mathcal{L} = \underbrace{-\lambda_L \langle w, (L_X, L_Y)\rangle}_{\text{maximize principal dependence}} + \underbrace{\lambda_R \langle w, (R_X, R_Y)\rangle}_{\text{minimize residual dependence}} + \underbrace{\lambda_{\text{rr}} \sum_{s\in\{X,Y\}} w_s \left[\frac{R_s}{L_s} - \tau_s\right]_+^2}_{\text{constrain relative residuals (with thresholds } \tau_s)} + \underbrace{\eta\,\Omega}_{\text{regularization}}, \tag{12}$$

where $[u]_+ = \max\{u, 0\}$, $\lambda_L, \lambda_R, \lambda_{\text{rr}}, \eta > 0$, Ω is a light stabilizer (e.g., whitening or $\|V^\top V - I\|_F^2$), $\langle , \rangle$ denotes Euclidean inner product on $\mathbb{R}^2$, $w = (w_X, w_Y)^\top$ are nonnegative branch weights, and $\tau_s > 0$ are per-branch tolerances.

For each evaluation, in addition to reporting lower and upper bounds, residuals, and gap metrics, uncertainty is quantified via bootstrap and permutation methods, yielding point estimates and confidence intervals:

Uncertainty via bootstrap. Report point estimates together with $(1-\alpha)$ confidence intervals for L_s, U_s, gap_s using nonparametric bootstrap.

Permutation calibration. Define $M_s = \frac{1}{2}\big(L_s + U_s\big)$. Under a permutation null, estimate (μ_0, σ_0) of M_s and report z-scores $z_s = (M_s - \mu_0)/\sigma_0$, lower-tailed p-values, and a calibrated index:

$$\text{CI}_s = \text{clip}\left(\frac{\mu_0 - M_s}{\mu_0 - L_s},\, 0,\, 1\right), \tag{13}$$

which maps null-like performance to 0 and the lower bound to 1.

3.5 Two Source Information Decomposition

3.5.1 Subspace Construction via Principal Angles

After obtaining the learned bottleneck representation $\mathbf{V}$, to directly support representation fusion, distillation, and model comparison, naturally, under the same metric, further quantify the deeper relationship between $\mathbf{V}$ and $R \in \{X, Y\}$. In general, given bottleneck representations from different models (e.g., $\mathbf{V_1}, \mathbf{V_2}$) information decomposition is employed to compute their shared and unique contributions with respect to R.

Consistent with CKA-based metric, given RFF bases $\mathbf{Q}_1, \mathbf{Q}_2$, compute the SVD of their overlap matrix $\mathbf{M} = \mathbf{Q}_1^\top \mathbf{Q}_2$:

$$\mathbf{M} = \tilde{\mathbf{U}}\,\mathbf{\Sigma}\,\tilde{\mathbf{V}}^\top, \qquad \mathbf{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_r),\;\; \sigma_k = \cos\theta_k.$$

Then, we adopt a threshold-free soft weighting. Let $t_k = \sigma_k^2 \in [0, 1]$ be the squared principal correlations. We choose a weighting function $\phi : [0, 1] \to [0, 1]$, such as $\phi(t) = t$. Write $\mathbf{W} = \text{diag}\big(\phi(t_1), \ldots, \phi(t_r)\big)$, define the following side–conservative and symmetric PSD operators:

$$\begin{aligned} P_{\cap,1}^{(\phi)} &= \mathbf{Q}_1\,\tilde{\mathbf{U}}\,\mathbf{W}\,\tilde{\mathbf{U}}^\top\mathbf{Q}_1^\top \qquad P_{1\setminus 2}^{(\phi)} = \mathbf{Q}_1\,\tilde{\mathbf{U}}\,(\mathbf{I} - \mathbf{W})\,\tilde{\mathbf{U}}^\top\mathbf{Q}_1^\top, \\ P_{\cap,2}^{(\phi)} &= \mathbf{Q}_2\,\tilde{\mathbf{V}}\,\mathbf{W}\,\tilde{\mathbf{V}}^\top\mathbf{Q}_2^\top \qquad P_{2\setminus 1}^{(\phi)} = \mathbf{Q}_2\,\tilde{\mathbf{V}}\,(\mathbf{I} - \mathbf{W})\,\tilde{\mathbf{V}}^\top\mathbf{Q}_2^\top. \end{aligned} \tag{14}$$

Theorem 2 (Positive semidefiniteness)**.** *If $\phi(t) \in [0, 1]$, then $P_{\cap,1}^{(\phi)}$, $P_{1\setminus 2}^{(\phi)}$, $P_{\cap,2}^{(\phi)}$, $P_{2\setminus 1}^{(\phi)}$ are PSD.*

Proof. For any $\mathbf{x}$, $\mathbf{x}^\top P_{\cap,1}^{(\phi)}\mathbf{x} = \|\mathbf{W}^{1/2}\tilde{\mathbf{U}}^\top\mathbf{Q}_1^\top\mathbf{x}\|_2^2 \geq 0$. The other cases are identical since $\mathbf{W} \succeq 0$ and $\mathbf{I} - \mathbf{W} \succeq 0$. □

Theorem 3 (Side-wise conservation)**.** *With the full SVD above ($\tilde{\mathbf{U}}, \tilde{\mathbf{V}} \in \mathbb{R}^{r\times r}$),*

$$P_{\cap,1}^{(\phi)} + P_{1\setminus 2}^{(\phi)} = \mathbf{Q}_1\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top\mathbf{Q}_1^\top, \qquad P_{\cap,2}^{(\phi)} + P_{2\setminus 1}^{(\phi)} = \mathbf{Q}_2\tilde{\mathbf{V}}\tilde{\mathbf{V}}^\top\mathbf{Q}_2^\top. \tag{15}$$

In particular, if $\tilde{\mathbf{U}}$ (resp. $\tilde{\mathbf{V}}$) is square orthogonal on the whole side, then the sum equals the side projector $\mathbf{Q}_1\mathbf{Q}_1^\top$ (resp. $\mathbf{Q}_2\mathbf{Q}_2^\top$).

3.5.2 Information Components w.r.t. a Reference Variable

Given a reference variable R with centered kernel $\mathbf{K}_R$, $\mathbf{K}_\Sigma = \frac{1}{2}(\mathbf{K}_1 + \mathbf{K}_2)$, $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ be the centering operator and $\widetilde{\mathbf{B}} = H\mathbf{B}H$, we evaluate information carried by each geometric component via $\widehat{I}_{\text{dep}}$:

$$I^{(\phi)}_{\text{red}|R} = \tfrac{1}{2}\Big[\widehat{I}_{\text{dep}}\big(\mathbf{K}_R,\ \widetilde{P^{(\phi)}_{\cap,1}\mathbf{K}_\Sigma P^{(\phi)}_{\cap,1}}\big) + \widehat{I}_{\text{dep}}\big(\mathbf{K}_R,\ \widetilde{P^{(\phi)}_{\cap,2}\mathbf{K}_\Sigma P^{(\phi)}_{\cap,2}}\big)\Big], \tag{16}$$

$$I^{(\phi)}_{1\setminus 2|R} = \widehat{I}_{\text{dep}}\big(\mathbf{K}_R,\ \widetilde{P^{(\phi)}_{1\setminus 2}\mathbf{K}_1 P^{(\phi)}_{1\setminus 2}}\big), \tag{17}$$

$$I^{(\phi)}_{2\setminus 1|R} = \widehat{I}_{\text{dep}}\big(\mathbf{K}_R,\ \widetilde{P^{(\phi)}_{2\setminus 1}\mathbf{K}_2 P^{(\phi)}_{2\setminus 1}}\big). \tag{18}$$

For $R = X$ (input-aligned), $I_{\text{red}|X}$ quantifies shared capture of input geometry and $I_{1\setminus 2|X}$ / $I_{2\setminus 1|X}$ quantify distinct input structures, for $R = Y$ (task-aligned), $I_{\text{red}|Y}$ measures label-relevant alignment shared by both sources, while $I_{1\setminus 2|Y}$ / $I_{2\setminus 1|Y}$ measure label-relevant alignment unique to each source.

Remark 2 (Non Additivity)**.** The operators P provide exact geometric cuts, then information can be measured via $\widehat{I}_{\text{dep}}$. Since $\widehat{I}_{\text{dep}}$ is a non additive log-ratio, we do not expect additivity, i.e., $I_{\text{red}} + I_{\text{uniq}} \neq I_{\text{total}}$.

3.6 Algorithm Flow

The complete algorithmic framework for learning information theoretic bottlenecks and performing two source decomposition is presented in Algorithm 1.

Algorithm 1 Kernel Alignment Bottleneck Learning and Information Decomposition

Input: Raw input X, labels Y, frozen embeddings $\mathbf{Z} = f(X) \in \mathbb{R}^{n\times a}$
Output: Bottleneck $\mathbf{V}$, information measures (L, R, U, gap) for X and Y, decomposition $(I_{\text{red}|X}, I_{\text{uniq}|X}, I_{\text{red}|Y}, I_{\text{uniq}|Y})$

`// Phase I: Data Preparation`

Compute fixed features $\mathbf{U}$ via Eq. 1, construct centered Gram matrices $\tilde{\mathbf{K}}_X, \tilde{\mathbf{K}}_Y, \tilde{\mathbf{K}}_Z$ using RBF kernel

`// Phase II: Bottleneck Learning`

Initialize learnable projection $T : \mathbf{Z} \to \mathbf{V}$ with $\dim(\mathbf{V}) = b \ll a$, **for** *each training iteration* **do**
  $\mathbf{V} \leftarrow T(\mathbf{Z})$ and compute $\tilde{\mathbf{K}}_V$
  Compute principal terms L_X, L_Y via Eq. 5
  Construct orthogonal residual kernel $\mathbf{G}_{\text{res}}$ via Eq. 6–7
  Compute residual terms R_X, R_Y via Eq. 8
  Compute training loss $\mathcal{L}$ via Eq. 12
  Update T via gradient descent

`// Phase III: Information Bounds and Uncertainty`

Compute upper bounds U_X, U_Y and gaps via Eq. 9, 10
Estimate confidence intervals via bootstrap resampling
Compute permutation based z-scores and calibrated indices via Eq. 13

`// Phase IV: Two Source Decomposition`

Input: Bottlenecks $\mathbf{V}_1, \mathbf{V}_2$ from two models
Compute principal angles via SVD: $\mathbf{M} = \mathbf{Q}_1^\top \mathbf{Q}_2 = \tilde{\mathbf{U}}\mathbf{\Sigma}\tilde{\mathbf{V}}^\top$
Construct geometric operators $P^{(\phi)}_{\cap,i}$ (Redundant) and $P^{(\phi)}_{i\setminus j}$ (unique) via Eq. 14
for *reference* $R \in \{X, Y\}$ **do**
  Compute redundant information $I^{(\phi)}_{\text{red}|R}$ via Eq. 18
  Compute unique information $I^{(\phi)}_{1\setminus 2|R}$, $I^{(\phi)}_{2\setminus 1|R}$ via Eq. 18

return Optimized $\mathbf{V}$, information bounds $(L_X, L_Y, R_X, R_Y, U_X, U_Y, \text{gap}_X, \text{gap}_Y)$, decomposition $(I_{\text{red}|X}, I_{1\setminus 2|X}, I_{2\setminus 1|X}, I_{\text{red}|Y}, I_{1\setminus 2|Y}, I_{2\setminus 1|Y})$

References

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

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMlR, 2019.

Robert Jenssen, Jose C Principe, Deniz Erdogmus, and Torbjørn Eltoft. The cauchy–schwarz divergence and parzen windowing: Connections to graph theory and mercer kernels. *Journal of the Franklin Institute*, 343(6):614–629, 2006.

Jose C Principe. *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media, 2010.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, pages 1177–1184, Vancouver, Canada, 2007.

Robert J Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 2009.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. *Advances in neural information processing systems*, 30, 2017.

Danica J Sutherland and Jeff Schneider. On the error of random fourier features. *arXiv preprint arXiv:1506.02785*, 2015.

A Appendix

A.1 Statistical Properties of $\widehat{I}_{\text{dep}}$

To validate $\widehat{I}_{\text{dep}}$ as a reliable statistical measure, we need additionally apply standard statistical theory to derive its consistency, bias, and asymptotic normality.

Population Quantities. Let $k_c(x, x')$ be the centered kernel, define the population-level quantities (targets of the V-statistics):

$$\mu_{KL} := \mathbb{E}[k_c(X, X')\ell_c(Z, Z')], \quad \mu_K := \mathbb{E}[k_c(X, X')^2], \quad \mu_L := \mathbb{E}[\ell_c(Z, Z')^2]$$

The true population target is $I^*_{\text{dep}} = -\log\left(\frac{\mu_{KL}^2}{\mu_K \mu_L}\right)$.

Lemma 1 (V-Statistic Consistency [Serfling, 2009])**.** *Assuming bounded and second-order integrable kernels, the normalized Frobenius inner products are V-statistics that converge almost surely to their population counterparts:*

$$\frac{1}{n^2}\langle \tilde{\mathbf{K}}, \tilde{\mathbf{L}} \rangle_F = \frac{1}{n^2}\sum_{i,j} k_c(X_i, X_j)\ell_c(Z_i, Z_j) \xrightarrow{a.s.} \mu_{KL}$$

$$\frac{1}{n^2}\langle \tilde{\mathbf{K}}, \tilde{\mathbf{K}} \rangle_F \xrightarrow{a.s.} \mu_K, \quad \frac{1}{n^2}\langle \tilde{\mathbf{L}}, \tilde{\mathbf{L}} \rangle_F \xrightarrow{a.s.} \mu_L$$

Proof. This follows directly from the strong law of large numbers for V-statistics [Serfling, 2009]. □

Lemma 2 (Consistency)**.** *Under the conditions of Lemma 1 and assuming non-degenerate kernels ($\mu_K, \mu_L > 0$), the estimator is strongly consistent:*

$$\widehat{I}_{dep}(\mathbf{K}, \mathbf{L}) \xrightarrow{a.s.} I^*_{dep}$$

Proof. This follows directly Lemma 1 and the continuous mapping theorem. □

Lemma 3 (Finite Sample Bias)**.** *Assume bounded and fourth-order integrable kernels. Let $I_n = \widehat{I}_{dep}(\mathbf{K}, \mathbf{L})$. The bias of the estimator is of order $O(n^{-1})$:*

$$\left|\mathbb{E}[I_n] - I^*_{dep}\right| \leq \frac{C}{n} + o\left(\frac{1}{n}\right)$$

Proof. Let $A_n = \frac{1}{n^2}\langle \tilde{K}, \tilde{L} \rangle_F$, $B_n = \frac{1}{n^2}\langle \tilde{K}, \tilde{K} \rangle_F$, $C_n = \frac{1}{n^2}\langle \tilde{L}, \tilde{L} \rangle_F$, $V_n = (A_n, B_n, C_n)^\top$, with population limits $\mu = (\mu_{KL}, \mu_K, \mu_L)^\top$. Each of A_n, B_n, C_n is a second order V-statistic, standard results give us:

$$\mathbb{E}[V_n] - \mu = O(n^{-1}).$$

Define $g(a, b, c) = -\log\big(a^2/(bc)\big)$ and $I_n = g(V_n)$, $I^* = g(\mu)$. A first order mean value expansion yields:

$$I_n - I^* = \nabla g(\mu)^\top (V_n - \mu) + R_n, \qquad \nabla g(\mu) = \big[-2/\mu_{KL},\, 1/\mu_K,\, 1/\mu_L\big]^\top,$$

where $\mathbb{E}|R_n| = o(n^{-1})$ under the stated moment and non–degeneracy assumptions. Taking expectations,

$$\left|\mathbb{E}[I_n] - I^*\right| \le \|\nabla g(\mu)\| \left\|\mathbb{E}[V_n] - \mu\right\| + o(n^{-1}) = O(n^{-1}).$$

□

Lemma 4 (Asymptotic Normality)**.** *Under the conditions of Lemma 3, the estimator is asymptotically normal:*

$$\sqrt{n}\left(\widehat{I}_{dep}(\mathbf{K}, \mathbf{L}) - I^*_{dep}\right) \xrightarrow{d} \mathcal{N}\left(0, \nabla g(\boldsymbol{\mu})^\mathsf{T} \boldsymbol{\Sigma} \nabla g(\boldsymbol{\mu})\right)$$

where $\boldsymbol{\mu} = [\mu_{KL}, \mu_K, \mu_L]^\mathsf{T}$, $\nabla g(\boldsymbol{\mu}) = [-2/\mu_{KL}, 1/\mu_K, 1/\mu_L]^\mathsf{T}$, *and* $\boldsymbol{\Sigma}$ *is the covariance matrix of the Hoeffding projections.*

Proof. Using the notation of Lemma 3, the vector of second order V-statistics satisfies joint CLT $\sqrt{n}\,(V_n - \mu) \Rightarrow \mathcal{N}(0, \Sigma)$. Since g is C^1 at μ, the multivariate Delta method yields $\sqrt{n}\,(g(V_n) - g(\mu)) \Rightarrow \mathcal{N}(0, \nabla g(\mu)^\top \Sigma \nabla g(\mu))$. □

A.2 Formal Justification of Geometric Residuals

To validate geometrically defined Eq. 8 is a valid and convergent statistical quantity, requires showing our sample based RFF projector Π converges to the correct population level operator.

Lemma 5 (RFF Projector Consistency)**.** *Let P_V be the true population level RKHS projection operator onto the subspace spanned by V. Let $\Pi = \Pi_{n,m}$ be the sample based projector from eq.* (6)*. If m is chosen appropriately (e.g., $m \gtrsim \sqrt{n}\log n$ [Rudi and Rosasco, 2017, Sutherland and Schneider, 2015]), then the sample projector satisfy:*

$$\|\Pi_{n,m} - P_V\|_{op} \xrightarrow{p} 0 \quad \text{as } n, m \to \infty.$$

Let $\varepsilon_n = \|(\Pi - P_V)\mathbf{K}_Z\|_F / \|\mathbf{K}_Z\|_F$. The conditions above imply $\varepsilon_n = o_p(1)$.

Proof. This result is a direct consequence of RFF kernel approximation [Rahimi and Recht, 2007] and kernel regression [Rudi and Rosasco, 2017]. □

Theorem 4 (Residual Estimator Convergence)**.** *Under the conditions of Lemma 5, the residual estimator $\widehat{I}_{dep}(\mathbf{K}_X, \tilde{\mathbf{G}}_{res})$ converges in probability to its population level target, the finite sample error is bounded:*

$$\left|\widehat{I}_{dep}(\mathbf{K}_X, \tilde{\mathbf{G}}_{res}) - I^*_{dep,\ pop\text{-}residual}\right| = O_p(\varepsilon_n + n^{-1/2})$$

Proof. The total error is a sum of the statistical (V-stat) error ($O_p(n^{-1/2})$) from **??** and the projector approximation error (ε_n) from Lemma 5. □

Proof. Recall, we have $\|\Pi_\lambda - P_V\|_{\mathrm{op}} \xrightarrow{p} 0$, hence $M_\lambda = I - \Pi_\lambda \to P_V^\perp$ and

$$\|\tilde{G}_{\mathrm{res}} - \widetilde{P_V^\perp K_Z P_V^\perp}\|_F \ \le\ C\,\|(\Pi_\lambda - P_V)K_Z\|_F \ =\ O_p(\varepsilon_n).$$

Let $F(B) := -\log\big(\langle \tilde{K}_X, \tilde{B}\rangle_F^2 / (\langle \tilde{K}_X, \tilde{K}_X\rangle_F \langle \tilde{B}, \tilde{B}\rangle_F)\big)$. On $\{\mu_X > 0, \mu_{\mathrm{res}} > 0, \mu_{X,\mathrm{res}} > 0\}$ the map F is C^1 in a neighborhood of $B_\star = \widetilde{P_V^\perp K_Z P_V^\perp}$, so by a mean–value bound

$$\left|F(\tilde{G}_{\mathrm{res}}) - F(B_\star)\right| = O_p(\varepsilon_n).$$

With $B_\star$ fixed, $F(B_\star)$ is a smooth function of order-2 V-statistics, hence by the joint CLT and the Delta method:

$$F_n(B_\star) - F(B_\star) = O_p(n^{-1/2}).$$

Combining the two displays gains $\left|\widehat{I}_{dep}(K_X, \tilde{G}_{\mathrm{res}}) - I^*_{dep,\mathrm{pop\text{-}res}}\right| = O_p(\varepsilon_n + n^{-1/2})$. □

A.3 Upper Bound Completeness

Proof. Let:

$$c := \mathrm{CKA}(A, B) = \frac{|\langle \widetilde{A}, \widetilde{B} \rangle_F|}{\|\widetilde{A}\|_F \|\widetilde{B}\|_F}, \quad c_1 := \mathrm{CKA}(A, B_\parallel), \quad c_2 := \mathrm{CKA}(A, B_\perp),$$

and set $\kappa = \|\widetilde{A}\|_F$, $a = \|\widetilde{B}_\parallel\|_F$, $b = \|\widetilde{B}_\perp\|_F$, $x = |\langle \widetilde{A}, \widetilde{B}_\parallel \rangle_F| = \kappa a\, c_1$, $y = |\langle \widetilde{A}, \widetilde{B}_\perp \rangle_F| = \kappa b\, c_2$. by triangle inequality:

$$c = \frac{x + y}{\kappa \|\widetilde{B}_\parallel + \widetilde{B}_\perp\|_F} \;\geq\; \frac{x + y}{\kappa(a + b)} = \frac{c_1 a + c_2 b}{a + b} \;\geq\; \min\{c_1, c_2\}.$$

Applying the monotone map $-2\log(\cdot)$ yield:

$$\widehat{I}_{dep}(A, B) \;\leq\; \max\{-2\log c_1,\, -2\log c_2\} = \max\{L_X, R_X\}.$$

Since $\max\{u, v\} \leq \log(e^u + e^v)$ for all $u, v \in \mathbb{R}$, we can obtain:

$$\widehat{I}_{dep}(A, B) \;\leq\; \log\big(e^{L_X} + e^{R_X}\big).$$

□