UNIFORM KERNEL PROBER

Anonymous authors
Paper under double-blind review

ABSTRACT

The ability to identify useful features or representations of the input data based on training data that achieves low prediction error on test data across multiple prediction tasks is considered the key to multitask learning success. In practice, however, one faces the issue of the choice of prediction tasks and the availability of test data from the chosen tasks while comparing the relative performance of different features. In this work, we develop a class of pseudometrics called Uniform Kernel Prober (UKP) for comparing features or representations learned by different statistical models such as neural networks when the downstream prediction tasks involve kernel ridge regression. The proposed pseudometric, UKP, between any two representations, provides a uniform measure of prediction error on test data corresponding to a general class of kernel ridge regression tasks for a given choice of a kernel without access to test data. Additionally, desired invariances in representations can be successfully captured by UKP only through the choice of the kernel function and the pseudometric can be efficiently estimated from n input data samples with $O(\frac{1}{\sqrt{n}})$ estimation error. We also experimentally demonstrate the ability of UKP to discriminate between different types of features or representations based on their generalization performance on downstream kernel ridge regression tasks.

1 INTRODUCTION

Model comparison is a classical problem in Statistics and Machine Learning (Burnham et al., 1998; Pfahringer et al., 2000; Spiegelhalter et al., 2002; Caruana & Niculescu-Mizil, 2006; Fernández-Delgado et al., 2014). This question has received tremendous attention from the scientific community, especially after the widespread adoption and implementation of modern general-purpose large-scale models such as deep neural networks (DNNs). Developing broadly applicable criteria for model comparison remains challenging due to variation in mathematical representations, no. of trainable parameters, and transparency (open vs. black-box access). In supervised learning, however, models can naturally be compared by differences in predictive performance, since this directly aligns with the goal of maximizing accuracy on the prediction task. It is now well understood that the key to success for training models with good generalization ability over multiple tasks (i.e. achieves low prediction error on test data across multiple prediction tasks) is directly correlated to the ability of models to identify useful features or representations of the input data based on training data (Bengio et al., 2013; LeCun et al., 2015; Maurer et al., 2016). Therefore, one can attempt to resolve the question of model comparison by considering metrics (more precisely, pseudometrics) on the space of features or representations, and there is extensive literature in this area (Laakso & Cottrell, 2000; Li et al., 2015; Morcos et al., 2018; Wang et al., 2018; Kornblith et al., 2019; Boix-Adsera et al., 2022).

An ideal pseudometric must be interpretable and efficiently computable based on a reasonably small amount of data samples. It must also be sensitive only to differences in features that will lead to differences in predictive performance, but be fairly insensitive to any other differences in features that do not affect predictive performance. Finally, it must be flexible enough to accommodate available prior knowledge about the class of prediction tasks that is of interest to the model users. However, most pseudometrics fall short of fulfilling this extensive set of desiderata. In this work, we develop a class of pseudometrics on the space of representations called Uniform Kernel Prober (UKP) that can be used to compare features or representations learned by any class of statistical models.

The proposed UKP pseudometric is motivated by the need for a distance measure over representations of differing dimensionalities that captures the ability of a model to generalize over a general and flexible class of prediction tasks, specifically, the class of kernel ridge regression-based tasks. Depending on the choice of the kernel, one can probe which models share "similar" features, with similarity being understood in the following sense: If the features or representations for a pair of models are similar, then, if they are both trained to perform kernel ridge regression tasks, their predictive performances will be close to each other. The UKP pseudometric is a unique distance measure over features or representations and is a useful contribution to the existing literature since it has the following desirable characteristics:

1. The proposed pseudometric offers a uniform guarantee of performance similarity for a wide range of regression functions, irrespective of whether the tasks are kernel ridge-regression or not. This is particularly beneficial when the prediction tasks align with models whose representations share similar characteristics with the kernel used to compute the UKP distance.
2. The pseudometric is adaptable to incorporate inductive biases that help identify models suited for specific tasks. A simple choice of the kernel parameter of the UKP distance can help us encode these inductive biases. For example, suppose we are interested in image classification tasks where the rotation and/or translation of the images should not affect the model prediction. In that case, we can encode this inductive bias into the pseudometric by choosing a rotationally and translationally invariant kernel, such as a Gaussian RBF kernel, as the kernel parameter for UKP. This results in the creation of two clusters: one for models with rotationally and translationally invariant features and another for models without such features.

   To the best of our knowledge, ours is the first pseudometric on the space of representations in the ML literature that can flexibly encode a wide range of inductive biases and treat them within a single framework.
3. UKP distance has a practical prediction-based interpretation in addition to usual mathematical interpretations of similarity or dissimilarity in terms of inner product or pseudometric.
4. Computation of the estimate of UKP distance only requires unlabelled data, i.e., data samples from the input domain, and therefore preserves labeled data for model training/fitting. Moreover, the computation of the estimate of UKP distance only requires black-box access to model representations, i.e., pairs of inputs and outputs to the model.
5. It is possible to design a statistically efficient estimator for the UKP distance based on a finite number (n) of samples from the input domain, that enjoys an estimation error rate of $n^{-1/2}$.
6. The UKP distance enables us to even compare representations that differ in their dimensionalities.

The paper is organized as follows. In Section 2, we formally define the UKP distance. In Section 3, we provide elegant and tractable characterizations of the UKP distance and prove that it satisfies all criteria of being a pseudometric. Then using Theorem 3, we also find the type of transformations under which the UKP distance remains invariant. We propose a statistical estimator of the UKP distance in Section 4. In Sections 4.1 and 4.2, we mathematically demonstrate its relationship to other pseudometrics used for model comparison and show that our proposed estimator converges to the true UKP distance as the sample size goes to infinity. Finally, in Section 5, we provide numerical experiments that validate our theory. Proofs of all lemmas, propositions and theorems are provided in Section A of the Appendix.

2 Problem Setup

Consider $\phi : \mathbb{R}^d \to \mathbb{R}^\ell$ and $\psi : \mathbb{R}^d \to \mathbb{R}^k$ to be two representation maps (with output dimensions l, k, respectively) that transform an input to its corresponding feature representation, typically obtained from a pair of trained/fitted models. Let Y be the random real-valued response corresponding to the input X generated from the nonparametric regression model $Y = \eta(X) + \epsilon$, where ϵ is mean-zero noise and $\eta(x) = \mathbb{E}(Y \mid X = x)$ is the population regression function of Y on X.

Let $K(\cdot,\cdot)$ be a positive definite, symmetric, bounded, and continuous "base" kernel function, mapping pairs of vectors in Euclidean spaces of different dimensions to real numbers. Common choices of radial kernels include the Gaussian RBF kernel $K_{RBF,h}(x,y) = \exp(-\frac{1}{2h}\|x-y\|_2^2)$ and the Laplace kernel $K_{Lap,h}(x,y) = \exp(-\frac{1}{2h}\|x-y\|_1)$, where $x, y \in \mathbb{R}^d$ for any $d \in \mathbb{N}$. By the Moore-Aronszajn Theorem (Aronszajn, 1950) and Lemma 4.33 of Steinwart & Christmann (2008), there exists a unique separable Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ of functions such that $K(\cdot,\cdot)$ is its unique reproducing kernel. Now, fix a single representation $\varphi : \mathbb{R}^d \to \mathbb{R}^{\text{out}}$. Define the corresponding "pullback kernel", $K_\varphi(\cdot,\cdot) \coloneqq K(\varphi(\cdot),\varphi(\cdot))$. By Theorem 5.7 of Paulsen & Raghupathi (2016), K_φ is positive-definite on $\mathbb{R}^d$ and is the (unique) reproducing kernel of the "pullback" RKHS $\mathcal{H}_\varphi \coloneqq \mathcal{H}(K \circ (\varphi \times \varphi))$. If we let $\mathcal{H}^{\text{out}}$ denote the RKHS associated with K when the domain is restricted to $\mathbb{R}^{\text{out}} \times \mathbb{R}^{\text{out}}$, then the $\mathcal{H}_\varphi$-norm (pullback RKHS norm) of any $f_\varphi \in \mathcal{H}_\varphi$ satisfies the minimal-norm characterization $\|f_\varphi\|_{\mathcal{H}_\varphi} = \min_{f \in \mathcal{H}^{\text{out}} : f \circ \varphi = f_\varphi} \|f\|_{\mathcal{H}^{\text{out}}}$.

For any two representations ϕ, ψ and for any $\lambda > 0$, let α_λ^ϕ and α_λ^ψ be the population kernel ridge regression estimators of the regression function η using their respective pullback kernels $K_\phi(\cdot,\cdot)$ and $K_\psi(\cdot,\cdot)$, defined as,

$$\alpha_\lambda^\phi = \arg\min_{f \in \mathcal{H}_\phi} \mathbb{E}\left[Y - f(X)\right]^2 + \lambda \|f\|_{\mathcal{H}_\phi}^2 \text{ and } \alpha_\lambda^\psi = \arg\min_{f \in \mathcal{H}_\psi} \mathbb{E}\left[Y - f(X)\right]^2 + \lambda \|f\|_{\mathcal{H}_\psi}^2, \tag{1}$$

respectively. Here, since the prediction loss is the squared error loss, α_λ^ϕ and α_λ^ψ depend on the distribution of Y only through the population regression function η. We suppress this dependence on η in the notation for convenience and clarity.

We now define the kernel ridge regression-based pseudometric between the two representations of the input ϕ and ψ, based on the difference between predictions for Y uniformly over all regression functions $\eta \in L^2(P_X)$ such that its $L^2(P_X)$ norm is bounded above by 1.

Definition 1. *For any $\lambda > 0$ and choice of kernel $K(\cdot,\cdot)$, the UKP (Uniform Kernel Prober) distance between representations $\phi(X)$ and $\psi(X)$ is defined as,*

$$d^{\text{UKP}}_{\lambda,K,\mathcal{L}^\infty}(\phi,\psi) \coloneqq \sup_{\|\eta\|_{L^2(P_X)} \le 1} \left(\mathbb{E}\left[\alpha_\lambda^\phi(X) - \alpha_\lambda^\psi(X)\right]^2\right)^{\frac{1}{2}},$$

where α_λ^ϕ and α_λ^ψ are defined in Equation1.

The reasoning behind the $\mathcal{L}^\infty$ in the notation for the UKP distance $d^{UKP}_{\lambda,K,\mathcal{L}^\infty}$ will be clear going forward, as we will demonstrate that it is actually an operator norm in an appropriate formal sense.

Remark 1. *In the main paper, we consider the UKP pseudometric w.r.t kernel ridge regression tasks only. Other types of regularization is also possible based on the manner in which the spectrum of the integral operators are regularized (Refer to Appendix A.2 for examples). We refer the readers to Definition 3 for the generalized definition of the UKP distance, denoted by $d^{\text{UKP}}_{g_\lambda,K,\mathcal{L}^\infty}$. All theoretical results in the main paper are proved for a general class of spectral regularizers in the Appendix.*

3 Properties of the UKP distance

Let $\mathfrak{I}_\phi : \mathcal{H}_\phi \to L^2(P_X), f \to f$ be the inclusion operator, which maps any $f \in \mathcal{H}_\phi$ to its representation $f \in L^2(P_X)$. Then the adjoint of the inclusion operator is given by $\mathfrak{I}_\phi^* : L^2(P_X) \to \mathcal{H}_\phi, f \to \int K_\phi(\cdot,x) f(x) dP_X(x)$. The inclusion operator $\mathfrak{I}_\psi$ and the corresponding adjoint operator $\mathfrak{I}_\psi^*$ can be analogously defined.

Let us define the covariance operators corresponding to the RKHS's $\mathcal{H}_\phi$ and $\mathcal{H}_\psi$ as

$$\Sigma_\phi \coloneqq \int K_\phi(\cdot,x) \otimes_{\mathcal{H}_\phi} K_\phi(\cdot,x) dP_X(x) = \int K(\phi(\cdot),\phi(x)) \otimes_{\mathcal{H}_\phi} K(\phi(\cdot),\phi(x)) dP_X(x)$$

and

$$\Sigma_\psi \coloneqq \int K_\psi(\cdot,x) \otimes_{\mathcal{H}_\psi} K_\psi(\cdot,x) dP_X(x) = \int K(\psi(\cdot),\psi(x)) \otimes_{\mathcal{H}_\psi} K(\psi(\cdot),\psi(x)) dP_X(x).$$

$\Sigma_\phi : \mathcal{H}_\phi \to \mathcal{H}_\phi$ and $\Sigma_\psi : \mathcal{H}_\psi \to \mathcal{H}_\psi$ are the unique operators that satisfy

$$\langle \Sigma_\phi f_1, g_1 \rangle_{\mathcal{H}_\phi} = \mathbb{E}\left[f_1(X) g_1(X)\right], \quad \langle \Sigma_\psi f_2, g_2 \rangle_{\mathcal{H}_\psi} = \mathbb{E}\left[f_2(X) g_2(X)\right],$$

where $f_1, g_1 \in \mathcal{H}_\phi$ and $f_2, g_2 \in \mathcal{H}_\psi$, respectively. In terms of inclusion operators, it can be easily shown that $\Sigma_\phi = \mathfrak{I}_\phi^* \mathfrak{I}_\phi$ and $\Sigma_\psi = \mathfrak{I}_\psi^* \mathfrak{I}_\psi$.

Let us define the integral operators corresponding to the RKHS's $\mathcal{H}_\phi$ and $\mathcal{H}_\psi$ as follows:

$$\mathcal{T}_\phi f \coloneqq \int K_\phi(\cdot, x) f(x) dP_X(x), \quad \mathcal{T}_\psi f \coloneqq \int K_\psi(\cdot, x) f(x) dP_X(x),$$

for any $f \in L^2(P_X)$. It is also easy to show that $\mathcal{T}_\phi = \mathfrak{I}_\phi \mathfrak{I}_\phi^*$ and $\mathcal{T}_\psi = \mathfrak{I}_\psi \mathfrak{I}_\psi^*$. The boundedness and continuity of the kernel K ensures that Σ_ϕ, Σ_ψ, $\mathcal{T}_\phi$ and $\mathcal{T}_\psi$ are all compact trace-class operators, which consequently ensures that they are also Hilbert-Schmidt operators. Further, each of Σ_ϕ, Σ_ψ, $\mathcal{T}_\phi$ and $\mathcal{T}_\psi$ are self-adjoint positive operators and therefore have a spectral representation (Reed & Simon, 1980, Theorems VI.16,VI.17). For any $\lambda > 0$, the regularized inverse covariance operators are defined as $\Sigma_\phi^{-\lambda} \coloneqq (\Sigma_\phi + \lambda I)^{-1}$ and $\Sigma_\psi^{-\lambda} \coloneqq (\Sigma_\psi + \lambda I)^{-1}$, while the corresponding square roots are defined as $\Sigma_\phi^{-\frac{\lambda}{2}} \coloneqq (\Sigma_\phi + \lambda I)^{-\frac{1}{2}}$ and $\Sigma_\psi^{-\frac{\lambda}{2}} \coloneqq (\Sigma_\psi + \lambda I)^{-\frac{1}{2}}$. Further, let us define $\widetilde{K}_\phi(x, y) \coloneqq \Sigma_\phi^{-\frac{\lambda}{2}} K_\phi(x, y)$ and $\widetilde{K}_\psi(x, y) \coloneqq \Sigma_\psi^{-\frac{\lambda}{2}} K_\psi(x, y)$. For any $p \geq 1$, we will use $\|\cdot\|_{\mathcal{L}^p(\mathcal{S})}$ to denote the p-Schatten norm of any operator mapping from its domain $\mathcal{S}$ into itself. In particular, for $p = 1, 2$ and ∞, the p-Schatten norm corresponds to the trace norm, Hilbert-Schmidt norm and the operator norm, respectively.

The UKP distance has the following characterization in terms of the integral operators, covariance operators and inclusion operators corresponding to the pullback kernels K_ϕ and K_ψ:

Theorem 1. *Assume that the base kernel K is defined on any Euclidean space and is positive definite, symmetric, bounded and continuous. Then, for any $\lambda > 0$, the the UKP distance $d^{\text{UKP}}_{\lambda, K, \mathcal{L}^\infty}(\phi, \psi)$ between representations $\phi(X)$ and $\psi(X)$ can be expressed as*

$$\begin{aligned} d^{\text{UKP}}_{\lambda, K, \mathcal{L}^\infty}(\phi, \psi) &= \left\| (\mathcal{T}_\phi + \lambda I)^{-1} \mathcal{T}_\phi - (\mathcal{T}_\psi + \lambda I)^{-1} \mathcal{T}_\psi \right\|_{\mathcal{L}^\infty(L^2(P_X))} \\ &= \left\| \mathfrak{I}_\phi (\Sigma_\phi + \lambda I)^{-1} \mathfrak{I}_\phi^* - \mathfrak{I}_\psi (\Sigma_\psi + \lambda I)^{-1} \mathfrak{I}_\psi^* \right\|_{\mathcal{L}^\infty(L^2(P_X))}. \end{aligned}$$

The proof is provided in Section A.4 of the Appendix as part of the proof of a generalized version of this theorem (Theorem 7). The above characterization shows that the UKP distance is the operator norm of the difference between a regularized/smoothed version of the integral operators corresponding to the pair of pullback RKHS's associated with the representations ϕ and ψ via the base kernel K. Using the monotonicity properties of p-Schatten norms (See Proposition 2.1 of Pfeiffer (2021)) we can develop a hierarchy of distances (pseudometrics), which we call generalized UKP distances, corresponding to the choice of the Schatten norm $\|\cdot\|_{\mathcal{L}^p(L^2(P_X))}$ for any $p \geq 1$, defined as follows:

Definition 2. *For any $\lambda > 0$, choice of kernel $K(\cdot, \cdot)$ and $p \geq 1$, the (λ, K, p)-UKP (Uniform Kernel Prober) distance between representations $\phi(X)$ and $\psi(X)$ is defined as,*

$$d^{\text{UKP}}_{\lambda, K, \mathcal{L}^p}(\phi, \psi) \coloneqq \left\| (\mathcal{T}_\phi + \lambda I)^{-1} \mathcal{T}_\phi - (\mathcal{T}_\psi + \lambda I)^{-1} \mathcal{T}_\psi \right\|_{\mathcal{L}^p(L^2(P_X))},$$

where $\mathcal{T}_\phi$ and $\mathcal{T}_\psi$ are the integral operators corresponding to the pullback RKHS's $\mathcal{H}_\phi$ and $\mathcal{H}_\psi$, respectively.

Next, we show that the (λ, K, p)-UKP distance indeed satisfies the axioms of a pseudometric for any valid choice of $\|\cdot\|_{\mathcal{L}^p(L^2(P_X))}$ corresponding to $\lambda > 0$ and $p \geq 1$. Of particular importance is the choice $p = 2$, which corresponds to the Hilbert-Schmidt norm, since it leads to a pseudometric which can be efficiently estimated using i.i.d samples from P_X. The question regarding whether it is possible to develop an estimator of $d^{UKP}_{\lambda, K, \mathcal{L}^\infty}(\phi, \psi)$ is still open because of the challenges involving operator norm estimation and the lack of inner product structure in such a scenario.

Theorem 2. *Assume that the setting of Theorem 1 holds true. Consider any three representations $\phi : \mathbb{R}^d \to \mathbb{R}^k$, $\psi : \mathbb{R}^d \to \mathbb{R}^l$ and $\varphi : \mathbb{R}^d \to \mathbb{R}^m$ for some $k, l, m \in \mathbb{N}$. Then, for any $1 \leq p \leq \infty$, we have that*

$$d^{\text{UKP}}_{\lambda, K, \mathcal{L}^\infty}(\phi, \psi) \leq d^{\text{UKP}}_{\lambda, K, \mathcal{L}^p}(\phi, \psi) \leq d^{\text{UKP}}_{\lambda, K, \mathcal{L}^1}(\phi, \psi) \tag{2}$$

Further, the $d^{\text{UKP}}_{\lambda,K,\mathcal{L}^p}(\phi,\psi)$ distance satisfies the following properties:

1. *(Positivity) $d^{\text{UKP}}_{\lambda,K,\mathcal{L}^p}(\phi,\phi)=0$,*
2. *(Non-negativity) $d^{\text{UKP}}_{\lambda,K,\mathcal{L}^p}(\phi,\psi)\geq 0$,*
3. *(Symmetricity) $d^{\text{UKP}}_{\lambda,K,\mathcal{L}^p}(\phi,\psi)=d^{\text{UKP}}_{\lambda,K,\mathcal{L}^p}(\psi,\phi)$,*
4. *(Triangle inequality) $d^{\text{UKP}}_{\lambda,K,\mathcal{L}^p}(\phi,\psi)\leq d^{\text{UKP}}_{\lambda,K,\mathcal{L}^p}(\phi,\varphi)+d^{\text{UKP}}_{\lambda,K,\mathcal{L}^p}(\varphi,\psi)$.*

Hence, $d^{\text{UKP}}_{\lambda,K,\mathcal{L}^p}(\phi,\psi)$ is a pseudometric over the space of all functions that maps $\mathbb{R}^d$ to some Euclidean space $\mathbb{R}^t$ for any $t\in\mathbb{N}$.

The proof is provided in Section A.5 of the Appendix as part of the proof of a more general result (Theorem 8). We now analyze the invariance properties of the pseudometric $d^{UKP}_{\lambda,K,\mathcal{L}^p}$ and identify the transformations of the representations ϕ and ψ that leave its value unchanged. Based on the following theorem, we can identify representations that UKP treats as equivalent in terms of prediction-based performance for a general collection of kernel ridge regression tasks corresponding to a particular kernel K. We achieve this goal by deriving an exact characterization of the representations that lead to $d^{UKP}_{\lambda,K,\mathcal{L}^p}=0$.

Theorem 3. *Assume that the setting of Theorem 1 holds true. Then, for any $p\geq 1$ and $\lambda>0$, given any two representations $\phi:\mathbb{R}^d\to\mathbb{R}^k$ and $\psi:\mathbb{R}^d\to\mathbb{R}^l$ we have that*

$$d^{\text{UKP}}_{\lambda,K,\mathcal{L}^p}(\phi,\psi)=0 \quad \textit{if and only if} \quad \mathcal{T}_\phi=\mathcal{T}_\psi. \tag{3}$$

Further, let $\mathcal{H}$ be the class of transformations under which the kernel K is invariant, i.e., $\mathcal{H}=\{h: K(\cdot,\cdot)=K(h(\cdot),h(\cdot))\ a.e.\ P_X\}$. Then, the UKP *distance $d^{\text{UKP}}_{\lambda,K,\mathcal{L}^p}(\phi,\psi)$ between representations $\phi(X)$ and $\psi(X)$ is invariant under the same class of transformations that the kernel K is invariant for, i.e., for any $h_1,h_2\in\mathcal{H}$,*

$$d^{\text{UKP}}_{\lambda,K,\mathcal{L}^p}(h_1\circ\phi,h_2\circ\psi)=d^{\text{UKP}}_{g_\lambda,K,\mathcal{L}^p}(\phi,\psi)$$

and if either h_1 or h_2 does not belong to $\mathcal{H}$,

$$d^{\text{UKP}}_{\lambda,K,\mathcal{L}^p}(h_1\circ\phi,h_2\circ\psi)\neq d^{\text{UKP}}_{\lambda,K,\mathcal{L}^p}(\phi,\psi).$$

Consequently, a necessary and sufficient condition for the UKP *distance $d^{\text{UKP}}_{g_\lambda,K,\mathcal{L}^p}(\phi,\psi)$ between representations $\phi(X)$ and $\psi(X)$ to be zero is that $K_\phi(\cdot,\cdot)=K_\psi(\cdot,\cdot)$ a.e. P_X.*

Additionally we can also provide a bound on the sensitivity of risk functional, uniformly over a class of kernel ridge regression tasks (Refer to Theorem 12 for more details).

Next, we show that the UKP distance $d^{UKP}_{\lambda,K,\mathcal{L}^2}(\phi,\psi)$ (corresponding to a Hilbert-Schmidt norm) can be expressed in terms of the trace operator, which will be essential for developing a statistical estimator of the pseudometric based on random samples from the input distribution P_X.

To do so, we define the cross-covariance operators $\Sigma_{\phi\psi}:\mathcal{H}_\psi\to\mathcal{H}_\phi$ and $\Sigma_{\psi\phi}:\mathcal{H}_\phi\to\mathcal{H}_\psi$ as follows:

$$\begin{aligned}\Sigma_{\phi\psi}&\coloneqq\int K_\phi(\cdot,x)\otimes_{\mathcal{L}^2(\mathcal{H}_\psi,\mathcal{H}_\phi)}K_\psi(\cdot,x)dP_X(x)\\&=\int K(\phi(\cdot),\phi(x))\otimes_{\mathcal{L}^2(\mathcal{H}_\psi,\mathcal{H}_\phi)}K(\psi(\cdot),\psi(x))dP_X(x)\end{aligned}$$

and

$$\begin{aligned}\Sigma_{\psi\phi}&\coloneqq\int K_\psi(\cdot,x)\otimes_{\mathcal{L}^2(\mathcal{H}_\phi,\mathcal{H}_\psi)}K_\phi(\cdot,x)dP_X(x)\\&=\int K(\psi(\cdot),\psi(x))\otimes_{\mathcal{L}^2(\mathcal{H}_\phi,\mathcal{H}_\psi)}K(\phi(\cdot),\phi(x))dP_X(x)=\Sigma^*_{\phi\psi}.\end{aligned}$$

Theorem 4. *For any $\lambda>0$, the squared* UKP *distance $d^{\text{UKP}}_{\lambda,K}(\phi,\psi)$ between representations $\phi(X)$ and $\psi(X)$ can be expressed as*

$$\left[d^{\text{UKP}}_{\lambda,K}(\phi,\psi)\right]^2=\text{Tr}\left(\Sigma_\phi^{-\lambda}\Sigma_\phi\Sigma_\phi^{-\lambda}\Sigma_\phi\right)+\text{Tr}\left(\Sigma_\psi^{-\lambda}\Sigma_\psi\Sigma_\psi^{-\lambda}\Sigma_\psi\right)-2\text{Tr}\left(\Sigma_\phi^{-\lambda}\Sigma_{\phi\psi}\Sigma_\psi^{-\lambda}\Sigma_{\psi\phi}\right).$$

The proof is provided in Section A.6 of the Appendix, as part of the proof of a more general result (Theorem 10).

4 STATISTICAL ESTIMATION OF $d^{UKP}_{\lambda,K,\mathcal{L}^2}$

In practice, when comparing the prediction-based utility of different representations, we consider the realistic scenario where one only has access to a random sample $X_1, \ldots, X_n \overset{i.i.d}{\sim} P_X$ and a statistical estimator of the proposed distance measure is required. In supervised learning settings, the goal is to allocate most of the data for training and model fitting while minimizing the amount of data used for diagnostics and exploratory analysis. Using the empirical covariance and cross-covariance operators $\hat{\Sigma}_\phi$, $\hat{\Sigma}_\psi$, $\hat{\Sigma}_{\phi\psi}$ and $\hat{\Sigma}_{\psi\phi} = \hat{\Sigma}^*_{\phi\psi}$ as plug-in estimators of Σ_ϕ, Σ_ψ, $\Sigma_{\phi\psi}$ and $\Sigma_{\psi\phi}$ in the trace operator based expression of $d^{\text{UKP}}_{\lambda,K}(\phi,\psi)$ as derived in Theorem 4, we arrive at the following V-statistic type estimator of $d^{\text{UKP}}_{\lambda,K}(\phi,\psi)$:

$$\hat{d}^{\text{UKP}}_{\lambda,K}(\phi,\psi) = \left[\operatorname{Tr}\left(\hat{\Sigma}_\phi^{-\lambda}\hat{\Sigma}_\phi\hat{\Sigma}_\phi^{-\lambda}\hat{\Sigma}_\phi\right) + \operatorname{Tr}\left(\hat{\Sigma}_\psi^{-\lambda}\hat{\Sigma}_\psi\hat{\Sigma}_\psi^{-\lambda}\hat{\Sigma}_\psi\right) - 2\operatorname{Tr}\left(\hat{\Sigma}_\phi^{-\lambda}\hat{\Sigma}_{\phi\psi}\hat{\Sigma}_\psi^{-\lambda}\hat{\Sigma}_{\psi\phi}\right)\right]^{\frac{1}{2}}, \tag{4}$$

where

$$\hat{\Sigma}_\phi = \frac{1}{n}\sum_{i=1}^n K_\phi(\cdot,X_i)\otimes_{\mathcal{H}_\phi} K_\phi(\cdot,X_i) = \frac{1}{n}\sum_{i=1}^n K(\phi(\cdot),\phi(X_i))\otimes_{\mathcal{H}_\phi} K(\phi(\cdot),\phi(X_i)),$$

$$\hat{\Sigma}_\psi = \frac{1}{n}\sum_{i=1}^n K_\psi(\cdot,X_i)\otimes_{\mathcal{H}_\psi} K_\psi(\cdot,X_i) = \frac{1}{n}\sum_{i=1}^n K(\psi(\cdot),\psi(X_i))\otimes_{\mathcal{H}_\psi} K(\psi(\cdot),\psi(X_i)),$$

$$\hat{\Sigma}_{\phi\psi} = \frac{1}{n}\sum_{i=1}^n K_\phi(\cdot,X_i)\otimes_{\mathcal{L}^2(\mathcal{H}_\psi,\mathcal{H}_\phi)} K_\psi(\cdot,X_i) = \frac{1}{n}\sum_{i=1}^n K(\phi(\cdot),\phi(X_i))\otimes_{\mathcal{L}^2(\mathcal{H}_\psi,\mathcal{H}_\phi)} K(\psi(\cdot),\psi(X_i)),$$

and

$$\begin{aligned}\hat{\Sigma}_{\psi\phi} =& \frac{1}{n}\sum_{i=1}^n K_\psi(\cdot,X_i)\otimes_{\mathcal{L}^2(\mathcal{H}_\phi,\mathcal{H}_\psi)} K_\phi(\cdot,X_i) = \frac{1}{n}\sum_{i=1}^n K(\psi(\cdot),\psi(X_i))\otimes_{\mathcal{L}^2(\mathcal{H}_\phi,\mathcal{H}_\psi)} K(\phi(\cdot),\phi(X_i))\\ =& \hat{\Sigma}^*_{\phi\psi}.\end{aligned}$$

It is an easy exercise to show that the V-statistic type estimator $\hat{d}^{\text{UKP}}_{\lambda,K}(\phi,\psi)$ can be expressed in terms of the number of input data points n, the chosen regularization parameter λ and the empirical Gram matrices $K_{n,\phi}$ and $K_{n,\psi}$ whose (i,j)-th elements are the kernel evaluations for the (i,j)-th input data pair (X_i, X_j), i.e., $(K_{n,\phi})_{ij} = K(\phi(X_i),\phi(X_j))$ and $(K_{n,\psi})_{ij} = K(\psi(X_i),\psi(X_j))$. If $\lambda = 0$, one is required to ensure the invertibility of $K_{n,\phi}$ and $K_{n,\psi}$.

Theorem 5. *For any $\lambda > 0$, the V-statistic type estimator $\hat{d}^{\text{UKP}}_{\lambda,K,\mathcal{L}^2}(\phi,\psi)$ of $d^{\text{UKP}}_{\lambda,K,\mathcal{L}^2}(\phi,\psi)$ between representations $\phi(X)$ and $\psi(X)$ can be expressed as*

$$\begin{aligned}&\hat{d}^{\text{UKP}}_{\lambda,K,\mathcal{L}^2}(\phi,\psi)\\ &= \big[\operatorname{Tr}\left(K_{n,\phi}(K_{n,\phi}+n\lambda I)^{-1}K_{n,\phi}(K_{n,\phi}+n\lambda I)^{-1}\right) + \operatorname{Tr}\left(K_{n,\psi}(K_{n,\psi}+n\lambda I)^{-1}K_{n,\psi}(K_{n,\psi}+n\lambda I)^{-1}\right)\\ &\quad - 2\operatorname{Tr}\left(K_{n,\phi}(K_{n,\phi}+n\lambda I)^{-1}K_{n,\psi}(K_{n,\psi}+n\lambda I)^{-1}\right)\big]^{\frac{1}{2}}.\end{aligned}$$

4.1 RELATION TO OTHER COMPARISON MEASURES

In this subsection, we discuss the relationship between the UKP distance and some popular distances between representations that are popularly used in Machine Learning. The UKP distance for the choice $p = 2$ (i.e. the Hilbert-Schmidt norm based $d^{UKP}_{\lambda,K,\mathcal{L}^2}$) is a generalization of the GULP distance, as proposed in Boix-Adsera et al. (2022), in the sense that, if we choose the kernel for the UKP to be the linear kernel $K_{lin}(x,y) = x^T y$, we exactly recover the GULP distance. Our proposed pseudometric $d^{UKP}_{\lambda,K,\mathcal{L}^p}$ provides the additional flexibility of choosing other kernel functions

(such as the Gaussian RBF kernel $K_{RBF,h}$ and the Laplace $K_{Lap,h}$) additional norms as well as additional regularization choices for understanding the relative difference between the generalization performance on different classes of kernel ridge regression-based prediction tasks.

Let $K_{n,\phi} = U_\phi \Lambda_{n,\phi} U_\phi^T$ and $K_{n,\psi} = U_\psi \Lambda_{n,\psi} U_\psi^T$ be the eigenvalue decompositions of $K_{n,\phi}$ and $K_{n,\psi}$, respectively. Here $\Lambda_{n,\phi} = \operatorname{diag}\left\{\mu_\phi^{(1)}, \ldots, \mu_\phi^{(n)}\right\}$ and $\Lambda_{n,\psi} = \operatorname{diag}\left\{\mu_\psi^{(1)}, \ldots, \mu_\psi^{(n)}\right\}$. Define $c_{\phi,\psi}^{(i),(j)} = \left(u_\phi^{(i)}\right)^T u_\psi^{(j)}$, as the inner product between the i-th eigenvector $u_\phi^{(i)}$ corresponding to the i-th eigenvalue $\mu_\phi^{(i)}$ of $K_{n,\phi}$ and j-th eigenvector $u_\psi^{(j)}$ corresponding to the j-th eigenvalue $\mu_\psi^{(i)}$ of $K_{n,\psi}$. In the following proposition, we express the V-statistic type estimator $\hat{d}^{UKP}_{\lambda,K,\mathcal{L}^2}(\phi,\psi)$ exclusively in terms of the inner products $c_{\phi,\psi}^{(i),(j)}$'s, the regularization parameter λ and the eigenvalues $\mu_\phi^{(i)}$'s and $\mu_\psi^{(j)}$'s, which is useful for understanding the effect of changing the regularization parameter λ on the estimate and its relation to other popular pseudometrics on the space of representations.

Proposition 1. *For any $\lambda > 0$, the V-statistic type estimator $\hat{d}^{\text{UKP}}_{\lambda,K,\mathcal{L}^2}(\phi,\psi)$ of $d^{\text{UKP}}_{\lambda,K,\mathcal{L}^2}(\phi,\psi)$ between representations $\phi(X)$ and $\psi(X)$ can be expressed as*

$$\hat{d}^{\text{UKP}}_{\lambda,K,\mathcal{L}^2}(\phi,\psi) = \left[\sum_{i=1}^{n}\left(\frac{\mu_\phi^{(i)}}{\mu_\phi^{(i)}+n\lambda}\right)^2 + \sum_{j=1}^{n}\left(\frac{\mu_\psi^{(j)}}{\mu_\psi^{(i)}+n\lambda}\right)^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\mu_\phi^{(i)}\mu_\psi^{(j)}}{\left(\mu_\phi^{(i)}+n\lambda\right)\left(\mu_\psi^{(j)}+n\lambda\right)}\left(c_{\phi,\psi}^{(i),(j)}\right)^2\right]^{\frac{1}{2}}.$$

The proof is straightforward, relying on the spectral decomposition of $K_{n,\phi}$ and $K_{n,\psi}$ and the properties of the trace operator, and is thus omitted.

The general kernelized version of the Ridge-CCA (Canonical Correlation Analysis) distance, introduced by Vinod (1976) and later discussed in M.Kuss & Graepel (2003), is defined as

$$\hat{d}^{\text{RCCA}}_{\lambda,K}(\phi,\psi) = \operatorname{Tr}\left(\hat{\Sigma}_\phi^{-\lambda}\hat{\Sigma}_{\phi\psi}\hat{\Sigma}_\psi^{-\lambda}\hat{\Sigma}_{\psi\phi}\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\mu_\phi^{(i)}\mu_\psi^{(j)}}{\left(\mu_\phi^{(i)}+n\lambda\right)\left(\mu_\psi^{(j)}+n\lambda\right)}\left(c_{\phi,\psi}^{(i),(j)}\right)^2.$$

However, the machine learning literature has largely focused on the original Ridge-CCA formulation with a linear kernel, as discussed in Kornblith et al. (2019). The classical CCA distance $\hat{d}^{\text{CCA}}$ can be derived from the kernelized Ridge-CCA distance $\hat{d}^{\text{RCCA}}_{\lambda,K}$ by selecting a linear kernel and setting $\lambda = 0$. From these definitions, it is clear that UKP is a distance measure on the Hilbert space of representations, while the kernelized Ridge-CCA serves as the corresponding inner product on the Hilbert space when the kernel and regularization parameter λ are the same for both.

Another related notion of distance, as proposed in Cristianini et al. (2001) and popularized by Kornblith et al. (2019), is known as CKA (Centered Kernel Alignment) and is defined as

$$\hat{d}^{\text{CKA}}_{K}(\phi,\psi) = \frac{\operatorname{Tr}\left(K_{n,\phi}H_nK_{n,\psi}H_n\right)}{\sqrt{\operatorname{Tr}\left(K_{n,\phi}H_nK_{n,\phi}H_n\right)\operatorname{Tr}\left(K_{n,\psi}H_nK_{n,\psi}H_n\right)}}$$

where $H_n = I_n - \frac{1}{n}1_n1_n^T$. We can equivalently express $\hat{d}^{\text{CKA}}_{K}(\phi,\psi)$ as

$$\hat{d}^{\text{CKA}}(\phi,\psi) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\mu_\phi^{(i)}\mu_\psi^{(j)}\left(c_{\phi,\psi}^{(i),(j)}\right)^2}{\sqrt{\sum_{i=1}^{n}\left(\mu_\phi^{(i)}\right)^2}\sqrt{\sum_{j=1}^{n}\left(\mu_\psi^{(j)}\right)^2}}.$$

If the kernelized Ridge-CCA distance is normalized by dividing it by the product of the norms of the pair of representations, taking the regularization parameter λ to $+\infty$ recovers the CKA measure $\hat{d}^{\text{CKA}}_{K}(\phi,\psi)$ in the limit. This can be shown by expressing $\hat{d}^{UKP}_{\lambda,K,\mathcal{L}^2}(\phi,\psi)$ and $\hat{d}^{\text{CKA}}_{K}(\phi,\psi)$ in terms of the eigenvalues and eigenvectors of the empirical Gram matrices $K_{n,\phi}$ and $K_{n,\psi}$ and then taking the limit as $\lambda \to +\infty$. The kernelized Ridge-CCA distance thus serves as a bridge between the CKA

measure, interpreted as a normalized inner product, and the UKP distance, understood as an unnormalized pseudometric in the space of representations. This connection implies a linear correlation between the two measures for sufficiently high value of the regularization parameter. While the CKA and kernelized Ridge-CCA measures naturally reflect similarity between representations via inner products, the UKP distance offers a broader perspective. Beyond functioning as a distance on the space of representations, it provides a relative measure of generalization performance uniformly across a wide range of prediction tasks involving kernel ridge regression—something other comparison measures fail to deliver. The UKP pseudometric efficiently compares learned representations by quantifying generalization similarity without task-specific training, leveraging pseudometric properties for meaningful and efficient assessment.

4.2 Finite sample convergence rate of $\hat{d}^{UKP}_{\lambda,K,\mathcal{L}^2}$

From a statistical estimation viewpoint, it is possible that the estimator $\hat{d}^{\text{UKP}}_{\lambda,K}$ converges to $d^{\text{UKP}}_{\lambda,K}$ as the number of data samples $X_1, \ldots, X_n$ from the input domain grows to infinity. In addition, we also provide a rate of convergence of the order of $O(\frac{1}{\sqrt{n}})$, which is a parametric rate of convergence. The following theorem, proved in Section A.7 of the Appendix, combines these two results and consequently illustrates the finite sample concentration of the estimator proposed in Equation equation 4 around the population $d^{UKP}_{\lambda,K,\mathcal{L}^2}$.

Theorem 6. *Let κ be an upper bound on the kernel function $K(\cdot,\cdot)$. Then, for any $\lambda > 0$ and $\delta > 0$, with probability atleast $1-\delta$, the V-statistic estimator $\hat{d}^{\text{UKP}}_{\lambda}(\phi,\psi)$ satisfies*

$$\left| \left(d^{\text{UKP}}_{\lambda,K,\mathcal{L}^2}(\phi,\psi)\right)^2 - \left(\hat{d}^{\text{UKP}}_{\lambda,K,\mathcal{L}^2}(\phi,\psi)\right)^2 \right| \leq \frac{8\kappa^3}{\lambda^3}\left[\frac{2\log(\frac{6}{\delta})}{n} + \sqrt{\frac{2\log(\frac{6}{\delta})}{n}}\right] + \frac{4\kappa^2}{\lambda^2}\left[\frac{2}{n} + \sqrt{\frac{2\log(\frac{6}{\delta})}{n}}\right].$$

For details regarding computational complexity, refer to Appendix A.10.

5 Experiments

In this section, we present experimental results that showcase the efficacy of the UKP distance in identifying similarities and differences between representations relevant to generalization performance on prediction tasks. For simplicity, we have considered $d^{UKP}_{\lambda,K,\mathcal{L}^2}$ as the UKP pseudometric and g_λ as the Tikhonov regularizer. Additional experiments, including model architecture details and training, are provided in the Appendix. All computations were performed on a single A100 GPU using Google Colab.

5.1 Ability of UKP to predict generalization performance by kernel ridge regression-based predictors

The UKP pseudometric gives a uniform bound on the difference in predictions generated by a pair of models, based on kernel ridge regression-based estimators that utilize the respective representations of the two models. It is a natural question to ask if this uniform or worst-case guarantee on the difference in prediction performance between representations is useful on a per-instance basis, i.e., given a specific kernel ridge regression task, whether the UKP distance is positively correlated with the generalization performance of different models. We consider 50 fully-connected neural networks with ReLU activation, each having uniform widths of 200, 400, 700, 800, or 900 and depths ranging from 1 to 10. These networks are trained on 60,000 28×28-pixel training images from the MNIST handwritten digits dataset (Deng, 2012) for 50 epochs. Representations are then extracted from the penultimate (final hidden) layer of each network, and the CCA, linear CKA (CKA with a linear kernel), GULP, and UKP distances are estimated for each pair of representations using 5,000 test images from the same dataset. We create synthetic kernel ridge regression tasks where we randomly sample 5000 images and randomly assign a standard Gaussian label to each image to create the synthetic label/target vector. We obtain the kernel ridge regression estimator for each representation with ridge penalty $\lambda \in \{10^{-2}, 1\}$ and Gaussian RBF kernel with bandwidth $\sigma \in \{10^{-1}, 1\}$. The empirical mean of the squared difference between predictions based on a pair of representations (say ϕ and ψ)

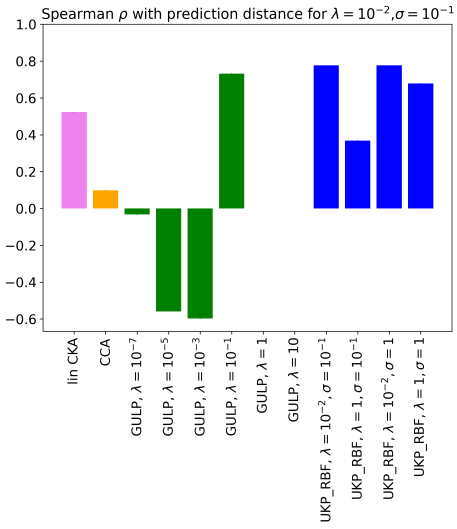


Figure 1: Generalization of kernel ridge regression-based predictors is strongly positively correlated with UKP distance values. We report the average correlation across 10 random synthetic kernel ridge regression tasks. Error bars are negligibly small and hence not visible.

is then computed using 5000 test images to estimate $err_{\phi,\psi} = \mathbb{E}_{X\sim P_X}\left[\alpha_\lambda^\phi(X) - \alpha_\lambda^\psi(X)\right]^2$, where α_λ^ϕ and α_λ^ψ are the kernel ridge regression based predictors. In Fig. 1, we plot the Spearman's ρ rank correlation coefficient between the $err_{\phi,\psi}$'s and the pairwise distances between the representations using CCA, linear CKA, GULP and UKP distances. For this particular regression task, we chose the synthetic ridge penalty to be $\lambda = 10^{-2}$ and used a Gaussian RBF kernel with $\sigma = 10^{-1}$. For the UKP distance, we use the Gaussian RBF kernel as the choice of kernel. We observe that the pairwise UKP distance is highly positively correlated with the collection of $err_{\phi,\psi}$'s, as evident from the large positive values of the blue bars, with the largest correlation being observed when the ridge penalty used in the UKP distance matches with the synthetic ridge penalty we chose, i.e., $\lambda = 10^{-2}$. In contrast, GULP distances exhibit inconsistent behavior across varying levels of regularization, while CCA and linear CKA distances show a significantly weaker positive correlation with generalization performance. As expected, due to the relationship between CKA and UKP discussed in Section 4.1, the CKA distance with a Gaussian RBF kernel performs comparably to UKP. Experiments with the remaining combinations of tuning parameters λ and σ are presented in Fig. 5 in Section B.2 of the Appendix, yielding qualitatively similar conclusions. We also discuss the ability of UKP to identify differences in architectures and inductive biases in Appendix B.1.

6 Conclusion and Future Work

This paper introduces the UKP pseudometric, a novel method for comparing model representations based on their predictive performance in kernel ridge regression tasks. It is shown to be easily interpretable, efficient, and capable of encoding inductive biases, supported by theoretical proofs and experimental validation. Therefore, the UKP pseudometric can serve as an useful and versatile exploratory tool for comparison of model representations, including representations learnt by black-box models such as neural networks, deep learning models and Large Language Models (LLMs). In our forthcoming work, we develop a conditional V-statistic type of estimator based on sample splitting and derive more sophisticated covergence guarantees with possibly better dependence on λ. Other research directions include using UKP for model selection, hyperparameter tuning, and enhancing its computational efficiency for large-scale models, such as deep neural networks, to better suit real-world applications.

REFERENCES

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Enric Boix-Adsera, Hannah Lawrence, George Stepaniants, and Philippe Rigollet. GULP: A prediction-based metric between representations. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 7115–7127. Curran Associates, Inc., 2022.

Kenneth P Burnham, David R Anderson, Kenneth P Burnham, and David R Anderson. *Practical use of the Information-Theoretic Approach*. Springer, 1998.

Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 161–168, 2006.

Nello Cristianini, John Shawe-Taylor, Andre Elisseeff, and Jaz Kandola. On kernel-target alignment. *Advances in Neural Information Processing Systems*, 14, 2001.

Li Deng. The MNIST database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1):3133–3181, 2014.

L. Lo Gerfo, L. Rosasco, F. Odone, E. De Vito, and A. Verri. Spectral algorithms for supervised learning. *Neural Computation*, 20(7):1873–1897, 07 2008. ISSN 0899-7667. doi: 10.1162/neco.2008.05-07-517. URL https://doi.org/10.1162/neco.2008.05-07-517.

Omar Hagrass, Bharath Sriperumbudur, and Bing Li. Spectral regularized kernel two-sample tests. *The Annals of Statistics*, 52(3):1076–1101, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Addison Howard, Eunbyung Park, and Wendy Kan. Imagenet object localization challenge. https://kaggle.com/competitions/imagenet-object-localization-challenge, 2018. Kaggle.

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

Aarre Laakso and Garrison Cottrell. Content and cluster analysis: Assessing representational similarity in neural systems. *Philosophical psychology*, 13(1):47–76, 2000.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? *arXiv preprint arXiv:1511.07543*, 2015.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

M.Kuss and T. Graepel. The geometry of kernel canonical correlation analysis. 2003.

Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. *Advances in Neural Information Processing Systems*, 31, 2018.

Vern I Paulsen and Mrinal Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*, volume 152. Cambridge university press, 2016.

Bernhard Pfahringer, Hilan Bensusan, and Christophe G Giraud-Carrier. Meta-learning by landmarking various learning algorithms. In *International Conference on Machine Learning*, pp. 743–750, 2000.

Paul Pfeiffer. On the stability of the area law for the entanglement entropy of the landau hamiltonian. *arXiv preprint arXiv:2102.07287*, 2021.

PyTorch. Models and pre-trained weights. `https://pytorch.org/vision/stable/models.html#classification`, 2024. Accessed: 2024-10-17.

Michael Reed and Barry Simon. *Methods of Modern Mathematical Physics: Functional Analysis*, volume 1. Gulf Professional Publishing, 1980.

Robert Schatten. *Norm ideals of completely continuous operators*, volume 27. Springer-Verlag, 2013.

David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

Bharath K Sriperumbudur and Nicholas Sterge. Approximate kernel PCA: Computational versus statistical trade-off. *The Annals of Statistics*, 50(5):2713–2736, 2022.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

Hrishikesh D Vinod. Canonical ridge and econometrics of joint production. *Journal of Econometrics*, 4(2):147–166, 1976.

Liwei Wang, Lunjia Hu, Jiayuan Gu, Zhiqiang Hu, Yue Wu, Kun He, and John Hopcroft. Towards understanding learning representations: To what extent do different neural networks learn the same representation. *Advances in Neural Information Processing Systems*, 31, 2018.

A Appendix: Proofs

In this appendix, we present the missing proofs of the paper.

A.1 Definitions and notations

For constants a and b, $a \lesssim b$ (*resp.* $a \gtrsim b$) denotes that there exists a positive constant c (*resp.* c') such that $a \leq cb$ (*resp.* $a \geq c'b$). $a \asymp b$ denotes that there exists positive constants c and c' such that $cb \leq a \leq c'b$. $[\ell]$ is used to denote $\{1, \ldots, \ell\}$. Let $\mathbf{1}_A$ denote the indicator function for the

Given a topological space $\mathcal{X}$, let $M_+^b(\mathcal{X})$ denote the space of all finite non-negative Borel measures on $\mathcal{X}$. We denote the space of bounded continuous functions defined on $\mathcal{X}$ by $C_b(\mathcal{X})$. For any $\mu \in M_+^b(\mathcal{X})$, let $L^r(\mathcal{X}, \mu)$ denote the Banach space of r-power $(r \geq 1)$ μ-integrable functions. For $f \in L^r(\mathcal{X}, \mu) =: L^r(\mu)$, we denote L^r-norm of f as $\|f\|_{L^r(\mu)} := (\int_{\mathcal{X}} |f|^r \, d\mu)^{1/r}$. $\mu^n := \mu \times \overset{n}{\ldots} \times \mu$ denotes the n-fold product measure. The equivalence class of the function f is defined as $[f]_\sim$ and consists of functions $g \in L^r(\mathcal{X}, \mu)$ such that $\|f - g\|_{L^r(\mu)} = 0$.

For any Hilbert space H, we denote the corresponding inner product and norm using $\langle \cdot, \cdot \rangle_H$ and $\|\cdot\|_H$, respectively. For any two abstract Hilbert spaces H_1 and H_2, let $\mathcal{L}(H_1, H_2)$ denote the space of bounded linear operators mapping from H_1 to H_2 and $\mathcal{L}^2(H_1, H_2)$ denote the space of Hilbert-Schmidt operators mapping from H_1 to H_2. For $M \in \mathcal{L}(H_1, H_2)$, its adjoint is denoted by M^*. $M \in \mathcal{L}(H) := \mathcal{L}(H, H)$ is called self-adjoint if $M^* = M$. For $M \in \mathcal{L}(H)$, $\mathrm{Tr}(M)$, $\|M\|_{\mathcal{L}^2(H)}$, and $\|M\|_{\mathcal{L}^\infty(H)}$ denote the trace, Hilbert-Schmidt and operator norms of M, respectively. For $x, y \in H$, $x \otimes_H y$ is an element of the tensor product space of $H \otimes H$ which can also be seen as an operator from $H \to H$ as $(x \otimes_H y)z = x\langle y, z\rangle_H$ for any $z \in H$. For any $M \in \mathcal{L}(H)$, we call it a positive definite (respectively, positive semi-definite) operator if $\langle f, Mf\rangle_H > 0$ (respectively, $\langle f, Mf\rangle_H \geq 0$) for any $f \in H$.

A.2 An introduction to spectral regularizers

Consider a spectral regularizer $g_\lambda : [0, \infty) \to \mathbb{R}$ which is a real-valued function compatible with a symmetric, bounded, positive definite and continuous kernel K with $\sup_x K(x, x) \leq \kappa$, in the sense that it satisfies the following regularity conditions, which are standard in the inverse problem and learning theory literature (Bauer et al., 2007; Hagrass et al., 2024)

$(\boldsymbol{A_1})$ $\sup_{x\in\Gamma} |x g_\lambda(x)| \leq C_1$;

$(\boldsymbol{A_2})$ $\sup_{x\in\Gamma} |\lambda g_\lambda(x)| \leq C_2$;

$(\boldsymbol{A_3})$ $\sup_{x\in\Gamma} |1 - x g_\lambda(x)|\, x^{2\varphi} \leq C_3 \lambda^{2\varphi}$ for $\varphi \in (0, \xi]$,

$(\boldsymbol{A_4})$ $g_\lambda(x) > 0$ for $x \in \Gamma \setminus \{0\}$ and $g_\lambda(0) \geq 0$,

$(\boldsymbol{A_5})$ $x \mapsto x g_\lambda(x)$ is an injective function for $x \in \Gamma$

where $\Gamma := [0, \kappa]$ and C_1, C_2 and C_3 are finite positive constants (all independent of λ). The constant ξ is usually termed as the *qualification* of g_λ and determines rates of convergence in the context of learning and hypothesis testing problems (Bauer et al., 2007; Hagrass et al., 2024). The intuition behind reasonable choices of g_λ is driven by the fact that we want $g_\lambda(\mathcal{B})\mathcal{B} = \mathcal{B} g_\lambda(\mathcal{B})$ to approximate the identity operator for small enough λ, which is ensured by Assumptions (A_1) and (A_3), since it ensures $\lim_{\lambda\to 0} x g_\lambda(x) \asymp 1$ (See for details Lemma A.20 of Hagrass et al. (2024) for details). Assumption (A_4) ensures that positive definite and positive semidefinite operators retain their respective definiteness properties after their eigenvalues are transformed/regularized by g_λ. Finally, Assumption (A_5) is a technical condition that is satisfied by popular spectral functions and its utility will be explained shortly,

Using functional calculus, given any self-adjoint operator $\mathcal{B} : \mathcal{H} \to \mathcal{H}$ with the spectral representation, $\mathcal{B} = \sum_i \tau_i \psi_i \otimes_H \psi_i$ with $(\tau_i, \psi_i)_i$ being the eigenvalues and eigenfunctions of $\mathcal{B}$, we can define the spectral regularization of $\mathcal{B}$ as

$$g_\lambda(\mathcal{B}) := \sum_{i\geq 1} g_\lambda(\tau_i)\,(\psi_i \otimes_{\mathcal{H}} \psi_i) + g_\lambda(0)\left(\boldsymbol{I} - \sum_{i\geq 1} \psi_i \otimes_{\mathcal{H}} \psi_i\right)$$

It is an easy exercise to show that if $g_\lambda(0) \neq 0$, then $g_\lambda(\mathcal{B})$ is invertible and self-adjoint. Further, if $g_\lambda(0) > 0$, then $g_\lambda(\mathcal{B})$ is positive definite. Finally, if $g_\lambda(x) > 0$ for all $x \in \Gamma = [0, \kappa]$ and satisfies Assumption (A_1) with $C_1 \leq 1$, then $[g_\lambda(\mathcal{B})]^{-1} - \mathcal{B}$ is self-adjoint and positive semi-definite. Additionally, if $C_1 < 1$, then $[g_\lambda(\mathcal{B})]^{-1} - \mathcal{B}$ is invertible, self-adjoint and positive definite. Finally, Assumption (A_5) guarantees that two self-adjoint positive semi-definite operators $\mathcal{B}_1$ and $\mathcal{B}_2$ are equal if and only if they share the same eigenfunctions and $\mathcal{B}_1 g_\lambda(\mathcal{B}_1) = \mathcal{B}_2 g_\lambda(\mathcal{B}_2)$.

A popular example of g_λ is $g_\lambda^{\mathrm{Tik}}(x) = \frac{1}{x+\lambda}$, yielding $g_\lambda^{\mathrm{Tik}}(\mathcal{B}) = (\mathcal{B} + \lambda \boldsymbol{I})^{-1}$, which is well known as the Tikhonov regularizer. Another example is the Showalter regularizer, $g_\lambda^{\mathrm{Sho}}(x) = \frac{1-e^{-x/\lambda}}{x}\mathbf{1}_{\{x\neq 0\}} + \frac{1}{\lambda}\mathbf{1}_{\{x=0\}}$. For both these regularizers, $g_\lambda(0) = \frac{1}{\lambda} > 0$. Note that the spectral cutoff regularizer is defined as $g_\lambda^{\mathrm{Cut}}(x) = \frac{1}{x}\mathbf{1}_{\{x\geq\lambda\}}$ satisfies Assumptions (A_1), (A_2) and (A_3) but $g_\lambda^{\mathrm{Cut}}(x) = 0$ for all $x < \lambda$ and $x g_\lambda^{\mathrm{Cut}}(x) = 1$ for all $x \geq \lambda$.

A.3 Preliminary results involving RKHS and RKHS-related operators

Lemma 1. *For any representations $\varphi : \mathbb{R}^d \to \mathbb{R}^m$ and $\vartheta : \mathbb{R}^d \to \mathbb{R}^t$, and positive definite, symmetric, bounded and continuous kernel K defined on any Euclidean space, let $K_\varphi(\cdot,\cdot) \coloneqq K(\varphi(\cdot),\varphi(\cdot))$ and $K_\vartheta(\cdot,\cdot) \coloneqq K(\vartheta(\cdot),\vartheta(\cdot))$ be the unique reproducing kernels corresponding to the "pullback" RKHS's $\mathcal{H}_\varphi \coloneqq \mathcal{H}(K \circ (\varphi \times \varphi))$ and $\mathcal{H}_\vartheta \coloneqq \mathcal{H}(K \circ (\vartheta \times \vartheta))$, respectively. Let $\mathfrak{I}_\varphi : \mathcal{H}_\varphi \to L^2(P_X), f \to f$ and $\mathfrak{I}_\vartheta : \mathcal{H}_\vartheta \to L^2(P_X), f \to f$ be the corresponding inclusion operators and $\mathfrak{I}_\varphi^* : L^2(P_X) \to \mathcal{H}_\varphi$ and $\mathfrak{I}_\vartheta^* : L^2(P_X) \to \vartheta$ be their corresponding adjoint operators. Define $\Sigma_\varphi = \mathfrak{I}_\varphi^*\mathfrak{I}_\varphi$ to be the covariance operator and $\mathcal{T}_\varphi = \mathfrak{I}_\varphi\mathfrak{I}_\varphi^*$ to be the integral operator corresponding to the RKHS $\mathcal{H}_\varphi$. Then, both $\Sigma_\varphi : \mathcal{H}_\varphi \to \mathcal{H}_\varphi$ and $\mathcal{T}_\varphi : L^2(P_X) \to L^2(P_X)$ are compact, self-adjoint and positive semi-definite operators. Further, define $\Sigma_{\varphi\vartheta} = \mathfrak{I}_\varphi^*\mathfrak{I}_\vartheta$ to be the cross-covariance operator mapping from $\mathcal{H}_\vartheta$ to $\mathcal{H}_\varphi$. Finally, consider a spectral regularizer g_λ that satisfies Assumptions (A_1), (A_2), (A_3) , (A_4) and (A_5). Then, both $g_\lambda(\Sigma_\varphi)$ and $g_\lambda(\mathcal{T}_\varphi)$ are also compact and self-adjoint operators.*

Then, we have that

(i) $\mathfrak{I}_\varphi^* : L^2(P_X) \to \mathcal{H}_\varphi, f \to \int K_\varphi(\cdot,x)f(x)dP_X(x)$

(ii) $(\Sigma_\varphi + \lambda I)^{-1}\mathfrak{I}_\varphi^* = \mathfrak{I}_\varphi^*(\mathcal{T}_\varphi + \lambda I)^{-1}$

(iii) $\mathfrak{I}_\varphi(\Sigma_\varphi + \lambda I)^{-1} = (\mathcal{T}_\varphi + \lambda I)^{-1}\mathfrak{I}_\varphi$

(iv) $\Sigma_\varphi(\Sigma_\varphi + \lambda I)^{-1} = (\Sigma_\varphi + \lambda I)^{-1}\Sigma_\varphi$

(v) $\mathcal{T}_\varphi(\mathcal{T}_\varphi + \lambda I)^{-1} = (\mathcal{T}_\varphi + \lambda I)^{-1}\mathcal{T}_\varphi$

(vi) $g_\lambda(\Sigma_\varphi)\mathfrak{I}_\varphi^* = \mathfrak{I}_\varphi^* g_\lambda(\mathcal{T}_\varphi)$

(vii) $\mathfrak{I}_\varphi g_\lambda(\Sigma_\varphi) = g_\lambda(\mathcal{T}_\varphi)\mathfrak{I}_\varphi$

(viii) $g_\lambda(\Sigma_\varphi)\Sigma_\varphi = \Sigma_\varphi g_\lambda(\Sigma_\varphi)$

(ix) $g_\lambda(\mathcal{T}_\varphi)\mathcal{T}_\varphi = \mathcal{T}_\varphi g_\lambda(\mathcal{T}_\varphi)$

(x) $\Sigma_{\varphi\vartheta} = \int K_\varphi(\cdot,x) \otimes_{\mathcal{L}^2(\mathcal{H}_\vartheta,\mathcal{H}_\varphi)} K_\vartheta(\cdot,x)dP_X(x) = \int K(\varphi(\cdot),\varphi(x)) \otimes_{\mathcal{L}^2(\mathcal{H}_\vartheta,\mathcal{H}_\varphi)} K(\vartheta(\cdot),\vartheta(x))dP_X(x)$

Proof. For any $f \in L^2(P_X)$ and $g \in \mathcal{H}_\varphi$, we have, by the definition of the adjoint of the inclusion operator $\mathfrak{I}_\varphi$,

$$\begin{aligned}
\left\langle \mathfrak{I}_\varphi^* f, g \right\rangle_{\mathcal{H}_\varphi} &= \langle f, \mathfrak{I}_\varphi g \rangle_{L^2(P_X)} \\
&= \int f(x)g(x)dP_X(x) \\
&= \int f(x) \left\langle K_\varphi(\cdot,x), g \right\rangle_{\mathcal{H}_\varphi} dP_X(x) \\
&= \left\langle \int K_\varphi(\cdot,x)f(x)dP_X(x), g \right\rangle_{\mathcal{H}_\varphi}
\end{aligned}$$

This proves (i).

Note that, $\mathfrak{I}_\varphi^*(\mathfrak{I}_\varphi\mathfrak{I}_\varphi^* + \lambda I) = (\mathfrak{I}_\varphi^*\mathfrak{I}_\varphi + \lambda I)\mathfrak{I}_\varphi^*$. By rearrangement, we obtain $(\mathfrak{I}_\varphi^*\mathfrak{I}_\varphi + \lambda I)^{-1}\mathfrak{I}_\varphi^* = \mathfrak{I}_\varphi^*(\mathfrak{I}_\varphi\mathfrak{I}_\varphi^* + \lambda I)^{-1}$, which proves (ii). Computing the adjoint of both sides of (ii) yields (iii).

Since K is a positive definite, symmetric, continuous and bounded kernel defined on a separable domain (the Euclidean space), the integral operator $\mathcal{T}_\varphi$ is a compact, self-adjoint and trace-class operator. Consequently, $\mathcal{T}_\varphi$ admits a spectral decomposition. Let $(\mu_i^\varphi, e_i^\varphi)_{i=1}^\infty$ be the eigenvalue-eigenfunction pairs corresponding to the spectral decomposition of $\mathcal{T}_\varphi$. Then, we have that

$$\mathcal{T}_\varphi = \sum_{i=1}^\infty \mu_i^\varphi \left(e_i^\varphi \otimes_{L^2(P_X)} e_i^\varphi\right)$$

Further, $(e_i^\varphi)_{i=1}^\infty$ constitutes an orthonormal basis of $L^2(P_X)$ and we must have that $\mu_i^\varphi > 0$ and $\lim_{i\to\infty} \mu_i^\varphi = 0$.

Using the fact that $\mathfrak{I}_\varphi^* \mathcal{T}_\varphi e_i^\varphi = \mathfrak{I}_\varphi^* \mathfrak{I}_\varphi (\mathfrak{I}_\varphi^* e_i^\varphi) = \Sigma_\varphi (\mathfrak{I}_\varphi^* e_i^\varphi)$ and $\left\| \mathfrak{I}_\varphi^* e_i^\varphi \right\|_{\mathcal{H}_\varphi} = (\mu_i^\varphi)^{\frac{1}{2}}$, we have that Σ_φ is also a compact, self-adjoint and trace-class operator which admits the following spectral decomposition

$$\Sigma_\varphi = \sum_{i=1}^\infty \mu_i^\varphi \left(\frac{\mathfrak{I}_\varphi^* e_i^\varphi}{\sqrt{\mu_i^\varphi}} \otimes_{\mathcal{H}_\varphi} \frac{\mathfrak{I}_\varphi^* e_i^\varphi}{\sqrt{\mu_i^\varphi}} \right)$$

Based on these spectral decompositions of $\mathcal{T}_\varphi$ and Σ_φ, we can readily derive that

$$\Sigma_\varphi (\Sigma_\varphi + \lambda I)^{-1} = \sum_{i=1}^\infty \frac{\mu_i^\varphi}{\mu_i^\varphi + \lambda} \left(\frac{\mathfrak{I}_\varphi^* e_i^\varphi}{\sqrt{\mu_i^\varphi}} \otimes_{\mathcal{H}_\varphi} \frac{\mathfrak{I}_\varphi^* e_i^\varphi}{\sqrt{\mu_i^\varphi}} \right) = (\Sigma_\varphi + \lambda I)^{-1} \Sigma_\varphi$$

and

$$\mathcal{T}_\varphi (\mathcal{T}_\varphi + \lambda I)^{-1} = \sum_{i=1}^\infty \frac{\mu_i^\varphi}{\mu_i^\varphi + \lambda} \left(e_i^\varphi \otimes_{L^2(P_X)} e_i^\varphi \right) = (\mathcal{T}_\varphi + \lambda I)^{-1} \mathcal{T}_\varphi$$

which completes the proof of (iv) and (v).

Further, the spectral decompositions of $g_\lambda(\mathcal{T}_\varphi)$ and $g_\lambda(\Sigma_\varphi)$ are given by

$$g_\lambda(\mathcal{T}_\varphi) = \sum_{i=1}^\infty g_\lambda(\mu_i^\varphi) \left(e_i^\varphi \otimes_{L^2(P_X)} e_i^\varphi \right) + g_\lambda(0) \left[I - \sum_{i=1}^\infty \left(e_i^\varphi \otimes_{L^2(P_X)} e_i^\varphi \right) \right]$$

and

$$g_\lambda(\Sigma_\varphi) = \sum_{i=1}^\infty g_\lambda(\mu_i^\varphi) \left(\frac{\mathfrak{I}_\varphi^* e_i^\varphi}{\sqrt{\mu_i^\varphi}} \otimes_{\mathcal{H}_\varphi} \frac{\mathfrak{I}_\varphi^* e_i^\varphi}{\sqrt{\mu_i^\varphi}} \right) + g_\lambda(0) \left[I - \sum_{i=1}^\infty \left(\frac{\mathfrak{I}_\varphi^* e_i^\varphi}{\sqrt{\mu_i^\varphi}} \otimes_{\mathcal{H}_\varphi} \frac{\mathfrak{I}_\varphi^* e_i^\varphi}{\sqrt{\mu_i^\varphi}} \right) \right]$$

which readily demonstrate the compactness and self-adjointness of these operators.

Using the above expressions of $g_\lambda(\mathcal{T}_\varphi)$ and $g_\lambda(\Sigma_\varphi)$, using the fact that $\mathfrak{I}_\varphi \mathfrak{I}_\varphi^* e_i^\varphi = \mu_i^\varphi e_i^\varphi$ and some elementary rearrangements, we can obtain (vi). Computing the adjoints of both sides of (vi), we obtain (vii). The commutativity properties (viii) and (ix) are also readily apparent from the spectral decompositions of $g_\lambda(\mathcal{T}_\varphi)$ and $g_\lambda(\Sigma_\varphi)$.

Finally, note that, for any $f \in \mathcal{H}_\vartheta$, we have that

$$\begin{aligned}
\Sigma_{\varphi\vartheta} =& \mathfrak{I}_\varphi^* \mathfrak{I}_\vartheta f \\
=& \int K_\varphi(\cdot, x) f(x) dP_X(x) \\
=& \int K_\varphi(\cdot, x) \left\langle K_\vartheta(\cdot, x), f \right\rangle_{\mathcal{H}_\vartheta} dP_X(x) \\
=& \left[\int K_\varphi(\cdot, x) \otimes_{\mathcal{L}^2(\mathcal{H}_\vartheta, \mathcal{H}_\varphi)} K_\vartheta(\cdot, x) dP_X(x) \right] f \\
=& \left[\int K(\varphi(\cdot), \varphi(x)) \otimes_{\mathcal{L}^2(\mathcal{H}_\vartheta, \mathcal{H}_\varphi)} K(\vartheta(\cdot), \vartheta(x)) dP_X(x) \right] f
\end{aligned}$$

which completes the proof of (x). □

Lemma 2. *For any representations $\varphi : \mathbb{R}^d \to \mathbb{R}^m$ and $\vartheta : \mathbb{R}^d \to \mathbb{R}^t$, and positive definite, symmetric, bounded and continuous kernel K defined on any Euclidean space, let $K_\varphi(\cdot,\cdot) \coloneqq K(\varphi(\cdot), \varphi(\cdot))$ and $K_\vartheta(\cdot,\cdot) \coloneqq K(\vartheta(\cdot), \vartheta(\cdot))$ be the unique reproducing kernels corresponding to the "pullback" RKHS's $\mathcal{H}_\varphi \coloneqq \mathcal{H}(K \circ (\varphi \times \varphi))$ and $\mathcal{H}_\vartheta \coloneqq \mathcal{H}(K \circ (\vartheta \times \vartheta))$, respectively. Given n i.i.d samples $\{X_i\}_{i=1}^n \sim P_X^n$, let $\mathcal{S}_\varphi : \mathcal{H}_\varphi \to \mathbb{R}^n, f \to \frac{1}{\sqrt{n}}(f(X_1), \ldots, f(X_n))^\top$ and $\mathcal{S}_\vartheta : \mathcal{H}_\vartheta \to \mathbb{R}^n, f \to \frac{1}{\sqrt{n}}(f(X_1), \ldots, f(X_n))^\top$ be the corresponding sampling operators, and let $\mathcal{S}_\varphi^* : \mathbb{R}^n \to \mathcal{H}_\varphi$ and $\mathcal{S}_\vartheta^* : \mathbb{R}^n \to \mathcal{H}_\vartheta$ be their adjoint operators. Define $\hat{\Sigma}_\varphi = \mathcal{S}_\varphi^* \mathcal{S}_\varphi$ to be the empirical covariance*

operator corresponding to the RKHS $\mathcal{H}_\varphi$ and $\hat{\Sigma}_{\varphi\vartheta} = \mathcal{S}_\varphi^* \mathcal{S}_\vartheta$ to be the empirical cross-covariance operator corresponding to the RKHS's mapping from $\mathcal{H}_\vartheta$ to $\mathcal{H}_\varphi$.

Further, define the empirical Gram matrices $K_{n,\varphi}$ and $K_{n,\vartheta}$ whose respective (i,j)-th elements are the kernel evaluations for the (i,j)-th input data pair (X_i, X_j), i.e., $(K_{n,\varphi})_{ij} = K(\phi(X_i), \phi(X_j))$ and $(K_{n,\vartheta})_{ij} = K(\vartheta(X_i), \vartheta(X_j))$.

Then, both $\hat{\Sigma}_\varphi : \mathcal{H}_\varphi \to \mathcal{H}_\varphi$ and $K_{n,\varphi} : \mathbb{R}^n \to \mathbb{R}^n$ are compact, self-adjoint and positive semi-definite operators. Further, we have that

(i) $\mathcal{S}_\varphi^* : \mathbb{R}^n \to \mathcal{H}_\varphi, \alpha = (\alpha_1, \ldots, \alpha_n)^\top \to \frac{1}{\sqrt{n}} \sum_{i=1}^n \alpha_i K_\varphi(\cdot, X_i)$

(ii) $\frac{1}{n} K_{n,\varphi} = \mathcal{S}_\varphi^* \mathcal{S}_\varphi$

(iii) $\hat{\Sigma}_\varphi = \frac{1}{n} \sum_{i=1}^n K_\varphi(\cdot, X_i) \otimes_{\mathcal{H}_\varphi} K_\varphi(\cdot, X_i) = \frac{1}{n} \sum_{i=1}^n K(\varphi(\cdot), \varphi(X_i)) \otimes_{\mathcal{H}_\varphi} K(\varphi(\cdot), \varphi(X_i))$

(iv) $\hat{\Sigma}_{\varphi\vartheta} = \frac{1}{n} \sum_{i=1}^n K_\varphi(\cdot, X_i) \otimes_{\mathcal{L}^2(\mathcal{H}_\vartheta, \mathcal{H}_\varphi)} K_\vartheta(\cdot, X_i) = \frac{1}{n} \sum_{i=1}^n K(\varphi(\cdot), \varphi(X_i)) \otimes_{\mathcal{L}^2(\mathcal{H}_\vartheta, \mathcal{H}_\varphi)} K(\vartheta(\cdot), \vartheta(X_i))$

(v) $\mathcal{S}_\varphi^* (\mathcal{S}_\varphi \mathcal{S}_\varphi^* + \lambda I)^{-1} = \mathcal{S}_\varphi^* (\frac{1}{n} K_{n,\vartheta} + \lambda I)^{-1} = (\hat{\Sigma}_\varphi + \lambda I)^{-1} \mathcal{S}_\varphi^* = (\mathcal{S}_\varphi^* \mathcal{S}_\varphi + \lambda I)^{-1} \mathcal{S}_\varphi^*$

(vi) $(\mathcal{S}_\varphi \mathcal{S}_\varphi^* + \lambda I)^{-1} \mathcal{S}_\varphi = (\frac{1}{n} K_{n,\vartheta} + \lambda I)^{-1} \mathcal{S}_\varphi = \mathcal{S}_\varphi (\hat{\Sigma}_\varphi + \lambda I)^{-1} = \mathcal{S}_\varphi (\mathcal{S}_\varphi^* \mathcal{S}_\varphi + \lambda I)^{-1}$

(vii) $g_\lambda(\hat{\Sigma}_\varphi) \mathfrak{I}_\varphi^* = \mathfrak{I}_\varphi^* g_\lambda(\frac{1}{n} K_{n,\varphi})$

(viii) $\mathfrak{I}_\varphi g_\lambda(\hat{\Sigma}_\varphi) = g_\lambda(\frac{1}{n} K_{n,\varphi}) \mathfrak{I}_\varphi$

(ix) $g_\lambda(\hat{\Sigma}_\varphi) \hat{\Sigma}_\varphi = \hat{\Sigma}_\varphi g_\lambda(\hat{\Sigma}_\varphi)$

(x) $g_\lambda(\frac{1}{n} K_{n,\varphi}) K_{n,\varphi} = K_{n,\varphi} g_\lambda(\frac{1}{n} K_{n,\varphi})$

(xi) For any operator $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathrm{Tr}(\mathcal{S}_\varphi^* A \mathcal{S}_\varphi) = \mathrm{Tr}(A \mathcal{S}_\varphi \mathcal{S}_\varphi^*)$

Proof. For any $f \in \mathcal{H}_\varphi$ and $\alpha \in \mathbb{R}^n$, we have, by definition of $\mathcal{S}_\varphi$ and adjoint operators, $\left\langle \mathcal{S}_\varphi^* \alpha, g \right\rangle_{\mathcal{H}_\varphi} = \langle \alpha, \mathcal{S}_\varphi g \rangle_2 = \frac{1}{\sqrt{n}} \sum_{i=1}^n \alpha_i g(X_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \alpha_i \langle K_\varphi(\cdot, X_i), g \rangle_{\mathcal{H}_\varphi} = \left\langle \frac{1}{\sqrt{n}} \sum_{i=1}^n \alpha_i K_\varphi(\cdot, X_i), g \right\rangle_{\mathcal{H}_\varphi}$. This completes the proof of (i)

The result in (ii) follows directly from the definitions of $K_{n,\varphi}$, $\mathcal{S}_\varphi$ and $\mathcal{S}_\varphi^*$. The rest of the proofs follow the same techniques used to prove Lemma 1. □

A.4 Expressing the UKP distance in terms of RKHS operators

Lemma 3. *Let Y be the random real-valued response corresponding to the input X generated from the nonparametric regression model $Y = \eta(X) + \epsilon$, where ϵ is mean-zero noise and $\eta(x) = \mathbb{E}(Y \mid X = x)$ is the population regression function of Y on X. For any $\lambda > 0$, consider a spectral regularizer g_λ that satisfies Assumptions (A_1), (A_2), (A_3) , (A_4) and (A_5) with $C_1 \leq 1$ and $g_\lambda(0) > 0$.*

Given a representation $\varphi : \mathbb{R}^d \to \mathbb{R}^m$, and positive definite, symmetric, bounded and continuous base kernel K defined on any Euclidean space, let $\alpha_{g_\lambda}^\varphi$ be the population kernel ridge regression estimator of the regression function η using the pullback kernel $K_\varphi(\cdot, \cdot) = K(\varphi(\cdot), \varphi(\cdot))$, defined as

$$\alpha_{g_\lambda}^\varphi = \underset{f \in \mathcal{H}_\varphi}{\arg\min}\, \mathbb{E}\left[Y - f(X)\right]^2 + \left\| (g_\lambda(\Sigma_\varphi)^{-1} - \Sigma_\varphi)^{\frac{1}{2}} f \right\|_{\mathcal{H}_\varphi}^2 \tag{5}$$

can be expressed as $\alpha_{g_\lambda}^\varphi = g_\lambda(\Sigma_\varphi) \mathfrak{I}_\varphi^* \eta \in \mathcal{H}_\varphi$. Consequently, as an element of $\mathcal{H}_\varphi$ embedded into $L^2(P_X)$, $\alpha_{g_\lambda}^\varphi$ can be expressed as

$$\mathfrak{I}_\varphi \alpha_{g_\lambda}^\varphi = \mathfrak{I}_\varphi g_\lambda(\Sigma_\varphi) \mathfrak{I}_\varphi^* \eta = \mathcal{T}_\varphi g_\lambda(\mathcal{T}_\varphi) \eta = g_\lambda(\mathcal{T}_\varphi) \mathcal{T}_\varphi \eta \in L^2(P_X). \tag{6}$$

In particular, when the chosen spectral regularizer is the Tikhonov regularizer $g_\lambda^{\mathrm{Tik}}(x) = \frac{1}{x+\lambda}$, we can express the solution to Equation 5 as

$$\alpha_\lambda^\varphi := \alpha_{g_\lambda^{\mathrm{Tik}}} = \operatorname*{arg\,min}_{f \in \mathcal{H}_\varphi} \mathbb{E}\left[Y - f(X)\right]^2 + \lambda \left\|f\right\|_{\mathcal{H}_\varphi}^2 = (\Sigma_\varphi + \lambda I)^{-1} \mathfrak{I}_\varphi^* \eta. \tag{7}$$

Consequently, as an element of $\mathcal{H}_\varphi$ embedded into $L^2(P_X)$, α_λ^φ can be expressed as

$$\mathfrak{I}_\varphi \alpha_\lambda^\varphi = \mathfrak{I}_\varphi (\Sigma_\varphi + \lambda I)^{-1} \mathfrak{I}_\varphi^* \eta = \mathcal{T}_\varphi (\mathcal{T}_\varphi + \lambda I)^{-1} \eta = (\mathcal{T}_\varphi + \lambda I)^{-1} \mathcal{T}_\varphi \eta \in L^2(P_X). \tag{8}$$

Proof. Consider a fixed population regression function $\eta(x) = \mathbb{E}(Y \mid X = x)$ corresponding to a fixed joint distribution P_{XY} of (X, Y) with marginal distribution of X as P_X. Note that, under the given conditions on the spectral regularizer g_λ and the kernel K, $g_\lambda(\Sigma_\varphi)$ is an invertible, self-adjoint and positive-definite operator, while $g_\lambda(\Sigma_\varphi)^{-1} - \Sigma_\varphi$ is a self-adjoint and positive semi-definite operator.

Now, for any $f \in \mathcal{H}_\varphi$, we have

$$\begin{aligned}
&\mathbb{E}\left[Y - f(X)\right]^2 + \left\|(g_\lambda(\Sigma_\varphi)^{-1} - \Sigma_\varphi)^{\frac{1}{2}} f\right\|_{\mathcal{H}_\varphi}^2 \\
=&\mathbb{E}\left[Y - \langle f, K_\varphi(\cdot, X)\rangle_{\mathcal{H}_\varphi}\right]^2 + \left\langle (g_\lambda(\Sigma_\varphi)^{-1} - \Sigma_\varphi)^{\frac{1}{2}} f, (g_\lambda(\Sigma_\varphi)^{-1} - \Sigma_\varphi)^{\frac{1}{2}} f \right\rangle_{\mathcal{H}_\varphi} \\
=&\mathbb{E}(Y^2) - 2\mathbb{E}\left[Y \langle f, K_\varphi(\cdot, X)\rangle_{\mathcal{H}_\varphi}\right] + \mathbb{E}\left[\langle f, K_\varphi(\cdot, X)\rangle_{\mathcal{H}_\varphi}^2\right] + \left\langle f, (g_\lambda(\Sigma_\varphi)^{-1} - \Sigma_\varphi) f \right\rangle_{\mathcal{H}_\varphi} \\
=&\mathbb{E}(Y^2) - 2\mathbb{E}\left[\eta(X) \langle f, K_\varphi(\cdot, X)\rangle_{\mathcal{H}_\varphi}\right] + \mathbb{E}\left\langle f, \left[K_\varphi(\cdot, X) \otimes_{\mathcal{H}_\varphi} K_\varphi(\cdot, X)\right] f \right\rangle_{\mathcal{H}_\varphi} \\
&+ \left\langle f, (g_\lambda(\Sigma_\varphi)^{-1} - \Sigma_\varphi) f \right\rangle_{\mathcal{H}_\varphi} \\
=&\mathbb{E}(Y^2) - 2\left\langle f, \mathfrak{I}_\varphi^* \eta \right\rangle_{\mathcal{H}_\varphi} + \langle f, \Sigma_\varphi f \rangle_{\mathcal{H}_\varphi} + \left\langle f, (g_\lambda(\Sigma_\varphi)^{-1} - \Sigma_\varphi) f \right\rangle_{\mathcal{H}_\varphi} \\
=&\mathbb{E}(Y^2) - 2\left\langle f, \mathfrak{I}_\varphi^* \eta \right\rangle_{\mathcal{H}_\varphi} + \left\langle f, (\Sigma_\varphi + g_\lambda(\Sigma_\varphi)^{-1} - \Sigma_\varphi) f \right\rangle_{\mathcal{H}_\varphi} \\
=&\mathbb{E}(Y^2) - 2\left\langle g_\lambda(\Sigma_\varphi)^{-\frac{1}{2}} f, g_\lambda(\Sigma_\varphi)^{\frac{1}{2}} \mathfrak{I}_\varphi^* \eta \right\rangle_{\mathcal{H}_\varphi} + \left\langle g_\lambda(\Sigma_\varphi)^{-\frac{1}{2}} f, g_\lambda(\Sigma_\varphi)^{-\frac{1}{2}} f \right\rangle_{\mathcal{H}_\varphi} \\
=&\mathbb{E}(Y^2) + \left\|g_\lambda(\Sigma_\varphi)^{-\frac{1}{2}} f - g_\lambda(\Sigma_\varphi)^{\frac{1}{2}} \mathfrak{I}_\varphi^* \eta\right\|_{\mathcal{H}_\varphi}^2 - \left\|g_\lambda(\Sigma_\varphi)^{\frac{1}{2}} \mathfrak{I}_\varphi^* \eta\right\|_{\mathcal{H}_\varphi}^2.
\end{aligned}$$

Therefore, the kernel ridge regression estimator of η using the representation $\varphi(X)$ and the pullback kernel K_φ is given by

$$\alpha_{g_\lambda}^\varphi = \operatorname*{arg\,min}_{f \in \mathcal{H}_\varphi} \mathbb{E}\left[Y - f(X)\right]^2 + \left\|(g_\lambda(\Sigma_\varphi)^{-1} - \Sigma_\varphi)^{\frac{1}{2}} f\right\|_{\mathcal{H}_\varphi}^2 = g_\lambda(\Sigma_\varphi) \mathfrak{I}_\varphi^* \eta \in \mathcal{H}_\varphi.$$

Using the inclusion operator $\mathfrak{I}_\varphi$, we can embed $\alpha_{g_\lambda}^\varphi$ in $L^2(P_X)$. Using this fact, together with Parts (vi) and (ix) of Lemma 1, we have that

$$\mathfrak{I}_\varphi \alpha_\lambda^\varphi = \mathfrak{I}_\varphi (\Sigma_\varphi + \lambda I)^{-1} \mathfrak{I}_\varphi^* \eta = \mathcal{T}_\varphi (\mathcal{T}_\varphi + \lambda I)^{-1} \eta = (\mathcal{T}_\varphi + \lambda I)^{-1} \mathcal{T}_\varphi \eta.$$

In particular, when the chosen spectral regularizer is the Tikhonov regularizer $g_\lambda^{\mathrm{Tik}}(x) = \frac{1}{x+\lambda}$, we can readily observe that $g_\lambda(\Sigma_\varphi) = (\Sigma_\varphi + \lambda I)^{-1}$ and $g_\lambda(\Sigma_\varphi)^{-1} - \Sigma_\varphi = \lambda I$. Substituting these expressions into the derivation of the closed form expression of $\alpha_{g_\lambda}^\varphi$ above leads to the closed form solution for $\alpha_\lambda^\varphi = \alpha_{g_\lambda^{\mathrm{Tik}}}^\varphi$ and yields the desired result, . □

Remark 2. *The population kernel ridge regression estimator of the regression function η using the pullback kernel $K_\varphi(\cdot, \cdot)$ given by $\alpha_{g_\lambda}^\varphi = g_\lambda(\Sigma_\varphi) \mathfrak{I}_\varphi^* \eta$ and its $L^2(P_X)$ embedding $\mathfrak{I}_\varphi^* \alpha_{g_\lambda}^\varphi = \mathcal{T}_\varphi g_\lambda(\mathcal{T}_\varphi) \eta = g_\lambda(\mathcal{T}_\varphi) \mathcal{T}_\varphi \eta$ is defined not only when the kernel regression problem in Equation 6 is well-posed $\Big($i.e. $g_\lambda(\Sigma_\varphi)^{-1} - \Sigma_\varphi$ is positive semi-definite, which is equivalent to convexity of the RKHS norm based penalty $\left\|(g_\lambda(\Sigma_\varphi)^{-1} - \Sigma_\varphi)^{\frac{1}{2}} f\right\|_{\mathcal{H}_\varphi}^2 \Big)$, but is well-defined in even more relaxed*

settings when the regularized kernel regression problem is ill-posed. In general, $\alpha^{\varphi}_{g_\lambda}$ and $\alpha^{\varphi}_{\lambda}$ can be interpreted as a low-pass spectral filter applied to the true regression function η that damp out the contribution of η along the "low energy" or less important eigenfunctions of Σ_φ (energy being a measure of the strength of alignment of η with a particular eigenfunction and is proportional to the magnitude of the corresponding eigenvalue) and retain/emphasize the contribution of η along the more the important eigenfunctions of Σ_φ (See Sections 3 and 4 of Gerfo et al. (2008) for a detailed discussion).

For any $\lambda > 0$, consider a spectral regularizer g_λ that satisfies Assumptions (A_1), (A_2), (A_3) , (A_4) and (A_5) with $C_1 \leq 1$ and $g_\lambda(0) > 0$. Given two representations $\phi : \mathbb{R}^d \to \mathbb{R}^k$ and $\psi : \mathbb{R}^d \to \mathbb{R}^l$, and positive definite, symmetric, bounded and continuous base kernel K defined on any Euclidean space, let $\alpha^{\phi}_{g_\lambda}$ and $\alpha^{\psi}_{g_\lambda}$ be the population kernel ridge regression estimators of the regression function η using their respective pullback kernels $K_\phi(\cdot,\cdot) = K(\phi(\cdot),\phi(\cdot))$ and $K_\psi(\cdot,\cdot) = K(\psi(\cdot),\psi(\cdot))$, defined as

$$\alpha^{\phi}_{g_\lambda} = \underset{f\in\mathcal{H}_\phi}{\arg\min}\, \mathbb{E}\left[Y - f(X)\right]^2 + \left\|(g_\lambda(\Sigma_\phi)^{-1} - \Sigma_\phi)^{\frac{1}{2}} f\right\|^2_{\mathcal{H}_\phi} \tag{9}$$

and

$$\alpha^{\psi}_{g_\lambda} = \underset{f\in\mathcal{H}_\psi}{\arg\min}\, \mathbb{E}\left[Y - f(X)\right]^2 + \left\|(g_\lambda(\Sigma_\psi)^{-1} - \Sigma_\psi)^{\frac{1}{2}} f\right\|^2_{\mathcal{H}_\psi} \tag{10}$$

In particular, when the chosen spectral regularizer is the Tikhonov regularizer $g^{\text{Tik}}_\lambda(x) = \frac{1}{x+\lambda}$, let α_λ and β_λ be the population kernel ridge regression estimators of the regression function η, given by

$$\alpha^{\phi}_{\lambda} = \underset{f\in\mathcal{H}_\phi}{\arg\min}\, \mathbb{E}\left[Y - f(X)\right]^2 + \lambda \left\|f\right\|^2_{\mathcal{H}_\phi} \tag{11}$$

and

$$\alpha^{\psi}_{\lambda} = \underset{f\in\mathcal{H}_\psi}{\arg\min}\, \mathbb{E}\left[Y - f(X)\right]^2 + \lambda \left\|f\right\|^2_{\mathcal{H}_\psi}, \tag{12}$$

respectively. The prediction loss being the squared error loss, $\alpha^{\phi}_{g_\lambda}$ and $\alpha^{\psi}_{g_\lambda}$ depend on the distribution of Y only through the population regression function η. We suppress this dependence on η in the notation for convenience and clarity.

We now define the kernel ridge regression-based pseudometric between the two representations of the input ϕ and ψ, based on the difference between predictions for Y uniformly over all regression functions $\eta \in L^2(P_X)$ such that its $L^2(P_X)$ norm is bounded above by 1.

Definition 3. *For any $\lambda > 0$, choice of kernel $K(\cdot,\cdot)$ and choice of spectral regularizer g_λ, the UKP (Uniform Kernel Prober) distance between representations $\phi(X)$ and $\psi(X)$ is defined as,*

$$d^{\text{UKP}}_{g_\lambda,K,\mathcal{L}^\infty}(\phi,\psi) \coloneqq \sup_{\|\eta\|_{L^2(P_X)}\leq 1} \left(\mathbb{E}\left[\alpha^{\phi}_{g_\lambda}(X) - \alpha^{\psi}_{g_\lambda}(X)\right]^2\right)^{\frac{1}{2}},$$

where $\alpha^{\phi}_{g_\lambda}$ and $\alpha^{\psi}_{g_\lambda}$ are defined in Equations 9 and 10, respectively.

Theorem 7. *For any $\lambda > 0$, consider a spectral regularizer g_λ that satisfies Assumptions (A_1), (A_2), (A_3) , (A_4) and (A_5) with $C_1 \leq 1$ and $g_\lambda(0) > 0$. Further, assume that the base kernel K is defined on any Euclidean space and is positive definite, symmetric, bounded and continuous. Then, the UKP distance $d^{\text{UKP}}_{g_\lambda,K,\mathcal{L}^\infty}(\phi,\psi)$ between representations $\phi(X)$ and $\psi(X)$ can be expressed as*

$$\begin{aligned} d^{\text{UKP}}_{g_\lambda,K,\mathcal{L}^\infty}(\phi,\psi) &= \|g_\lambda(\mathcal{T}_\phi)\mathcal{T}_\phi - g_\lambda(\mathcal{T}_\psi)\mathcal{T}_\psi\|_{\mathcal{L}^\infty(L^2(P_X))} \\ &= \|\mathcal{T}_\phi g_\lambda(\mathcal{T}_\phi) - \mathcal{T}_\psi g_\lambda(\mathcal{T}_\psi)\|_{\mathcal{L}^\infty(L^2(P_X))} \\ &= \left\|\mathfrak{I}_\phi g_\lambda(\Sigma_\phi)\mathfrak{I}^*_\phi - \mathfrak{I}_\psi g_\lambda(\Sigma_\psi)\mathfrak{I}^*_\psi\right\|_{\mathcal{L}^\infty(L^2(P_X))}. \end{aligned}$$

In particular, when the chosen spectral regularizer is the Tikhonov regularizer $g^{\text{Tik}}_\lambda(x) = \frac{1}{x+\lambda}$, the UKP distance $d^{\text{UKP}}_{\lambda,K,\mathcal{L}^\infty}(\phi,\psi) \coloneqq d^{\text{UKP}}_{g^{\text{Tik}}_\lambda,K,\mathcal{L}^\infty}(\phi,\psi)$ can be expressed as

$$\begin{aligned} d^{\text{UKP}}_{\lambda,K,\mathcal{L}^\infty}(\phi,\psi) &= \left\|(\mathcal{T}_\phi + \lambda I)^{-1}\mathcal{T}_\phi - (\mathcal{T}_\psi + \lambda I)^{-1}\mathcal{T}_\psi\right\|_{\mathcal{L}^\infty(L^2(P_X))} \\ &= \left\|\mathcal{T}_\phi(\mathcal{T}_\phi + \lambda I)^{-1} - \mathcal{T}_\psi(\mathcal{T}_\psi + \lambda I)^{-1}\right\|_{\mathcal{L}^\infty(L^2(P_X))} \\ &. = \left\|\mathfrak{I}_\phi(\Sigma_\phi + \lambda I)^{-1}\mathfrak{I}^*_\phi - \mathfrak{I}_\psi(\Sigma_\psi + \lambda I)^{-1}\mathfrak{I}^*_\psi\right\|_{\mathcal{L}^\infty(L^2(P_X))}. \end{aligned}$$

Proof. Under the conditions imposed on the spectral regularizer g_λ and the base kernel K, the population kernel regression estimators $\alpha^\phi_{g_\lambda}$ and $\alpha^\psi_{g_\lambda}$ as defined in Equations 9 and 10 and used in Definition 3 are explicitly given by $\alpha^\phi_{g_\lambda} = g_\lambda(\Sigma_\phi)\mathfrak{I}^*_\phi\eta$ and $\alpha^\psi_{g_\lambda} = g_\lambda(\Sigma_\psi)\mathfrak{I}^*_\psi\eta$, with their corresponding $L^2(P_X)$ embeddings being $\mathfrak{I}^*_\phi\alpha^\phi_{g_\lambda} = \mathcal{T}_\phi g_\lambda(\mathcal{T}_\phi)\eta = g_\lambda(\mathcal{T}_\phi)\mathcal{T}_\phi\eta$ and $\mathfrak{I}^*_\psi\alpha^\psi_{g_\lambda} = \mathcal{T}_\psi g_\lambda(\mathcal{T}_\psi)\eta = g_\lambda(\mathcal{T}_\psi)\mathcal{T}_\psi\eta$, respectively. Consequently, for any fixed $\eta = \mathbb{E}(Y \mid X) \in L^2(P_X)$, we have that

$$\begin{aligned} E\left[\alpha^\phi_{g_\lambda}(X) - \alpha^\psi_{g_\lambda}(X)\right]^2 &= \int \left[\alpha^\phi_{g_\lambda}(X) - \alpha^\psi_{g_\lambda}(X)\right]^2 dP_X(x) \\ &= \left\|\mathfrak{I}_\phi\alpha^\phi_{g_\lambda} - \mathfrak{I}_\psi\alpha^\psi_{g_\lambda}\right\|^2_{L^2(P_X)} \\ &= \left\|g_\lambda(\mathcal{T}_\phi)\mathcal{T}_\phi\eta - g_\lambda(\mathcal{T}_\psi)\mathcal{T}_\psi\eta\right\|^2_{L^2(P_X)} \\ &= \left\|\left[g_\lambda(\mathcal{T}_\phi)\mathcal{T}_\phi - g_\lambda(\mathcal{T}_\psi)\mathcal{T}_\psi\right]\eta\right\|^2_{L^2(P_X)}. \end{aligned}$$

Finally, using the definition of the operator norm of $[g_\lambda(\mathcal{T}_\phi)\mathcal{T}_\phi - g_\lambda(\mathcal{T}_\psi)\mathcal{T}_\psi]$ together with Parts (vi) and (ix) of Lemma 1, we obtain that

$$\begin{aligned} d^{UKP}_{g_\lambda,K,\mathcal{L}^\infty}(\phi,\psi) &= \sup_{\|\eta\|_{L^2(P_X)}\leq 1} \left(\mathbb{E}\left[\alpha_\lambda(X) - \beta_\lambda(X)\right]^2\right)^{\frac{1}{2}} \\ &= \sup_{\|\eta\|_{L^2(P_X)}\leq 1} \left\|\left[g_\lambda(\mathcal{T}_\phi)\mathcal{T}_\phi - g_\lambda(\mathcal{T}_\psi)\mathcal{T}_\psi\right]\eta\right\|_{L^2(P_X)} \\ &= \left\|g_\lambda(\mathcal{T}_\phi)\mathcal{T}_\phi - g_\lambda(\mathcal{T}_\psi)\mathcal{T}_\psi\right\|_{\mathcal{L}^\infty(L^2(P_X))} \\ &= \left\|\mathcal{T}_\phi g_\lambda(\mathcal{T}_\phi) - \mathcal{T}_\psi g_\lambda(\mathcal{T}_\psi)\right\|_{\mathcal{L}^\infty(L^2(P_X))} \\ &= \left\|\mathfrak{I}_\phi g_\lambda(\Sigma_\phi)\mathfrak{I}^*_\phi - \mathfrak{I}_\psi g_\lambda(\Sigma_\psi)\mathfrak{I}^*_\psi\right\|_{\mathcal{L}^\infty(L^2(P_X))}. \end{aligned}$$

In particular, when the chosen spectral regularizer is the Tikhonov regularizer $g^{\mathrm{Tik}}_\lambda(x) = \frac{1}{x+\lambda}$, we obtain the required result for $d^{UKP}_{\lambda,K,\mathcal{L}^\infty} = d^{UKP}_{g^{\mathrm{Tik}}_\lambda,K,\mathcal{L}^\infty}$.

□

Remark 3. *A very specific case of our proposed pseudometric $d^{\mathrm{UKP}}_{g_\lambda,K,\mathcal{L}^p}$, named the GULP distance, was analyzed in Boix-Adsera et al. (2022). However, the authors of Boix-Adsera et al. (2022) inaccurately derived GULP to correspond to a Hilbert-Schmidt norm, when it should actually correspond to an operator norm. This stems from the fact that, in the proof of Lemma 1 in Section A.1 of Boix-Adsera et al. (2022) (Page 14 of both the NeurIPS version Boix-Adsera et al. (2022) and the ArXiv version), the squared GULP distance is expressed as the supremum of an expectation (as we do in Definition 1) but they erroneously interchange the supremum and the expectation. By Jensen's inequality, the resulting quantity after this interchange can be shown to be greater than or equal to the definition of GULP distance. We claim that the distance actually analyzed in Boix-Adsera et al. (2022) corresponds to making the specific choices of $p = 2$, spectral regularizer $g_\lambda = g^{\mathrm{Tik}}_\lambda$ and the linear base kernel $K = K_{\mathrm{lin}}(x,y) = x^T y$ in our proposed UKP distance $d^{\mathrm{UKP}}_{g_\lambda,K,\mathcal{L}^p}$.*

A.5 Properties of the UKP distance

For any $p \geq 1$, we will use $\|\cdot\|_{\mathcal{L}^p(\mathcal{S})}$ to denote the p-Schatten norm of any operator mapping from its domain $\mathcal{S}$ into itself. In particular, for $p = 1, 2$ and ∞, the p-Schatten norm corresponds to the trace norm, Hilbert-Schmidt norm and the operator norm, respectively.

Using the monotonicity properties of p-Schatten norms (See Proposition 2.1 of Pfeiffer (2021)) we can develop a hierarchy of distances (pseudometrics), which we call generalized UKP distances, corresponding to the choice of the Schatten norm $\|\cdot\|_{\mathcal{L}^p(L^2(P_X))}$ for any $p \geq 1$, defined as follows:

Definition 4. *For any $\lambda > 0$, choice of kernel $K(\cdot,\cdot)$,choice of spectral regularizer g_λ and $p \geq 1$, the (g_λ, K, p)-UKP (Uniform Kernel Prober) distance between representations $\phi(X)$ and $\psi(X)$ is defined as,*

$$d^{\mathrm{UKP}}_{g_\lambda,K,\mathcal{L}^p}(\phi,\psi) \coloneqq \left\|g_\lambda(\mathcal{T}_\phi)\mathcal{T}_\phi - g_\lambda(\mathcal{T}_\psi)\mathcal{T}_\psi\right\|_{\mathcal{L}^p(L^2(P_X))},$$

where $\mathcal{T}_\phi$ and $\mathcal{T}_\psi$ are the integral operators corresponding to the pullback RKHS's $\mathcal{H}_\phi$ and $\mathcal{H}_\psi$, respectively.

Further, the following theorem also serves to show that the (g_λ, K, p)-UKP distance satisfies the axioms of a pseudometric for any valid choice of $\|\cdot\|_{\mathcal{L}^p(L^2(P_X))}$ and spectral regularizer g_λ.

Of particular importance is the choice $p = 2$, which corresponds to the Hilbert-Schmidt norm, since it leads to a pseudometric which can be efficiently estimated using i.i.d samples from P_X.

Theorem 8. *Assume that the setting of Theorem 7 holds true. Then, for any $1 \leq p \leq \infty$, we have that*

$$d^{\text{UKP}}_{g_\lambda,K,\mathcal{L}^\infty}(\phi,\psi) \leq d^{\text{UKP}}_{g_\lambda,K,\mathcal{L}^p}(\phi,\psi) \leq d^{\text{UKP}}_{g_\lambda,K,\mathcal{L}^1}(\phi,\psi) \tag{13}$$

Further, the $d^{\text{UKP}}_{g_\lambda,K,\mathcal{L}^p}(\phi,\psi)$ distance satisfies the following properties:

1. *For any function $\phi : \mathbb{R}^d \to \mathbb{R}^k$ for some $k \in \mathbb{N}$, $d^{\text{UKP}}_{g_\lambda,K,\mathcal{L}^p}(\phi,\phi) = 0$,*
2. *(Non-negativity) For any two functions $\phi : \mathbb{R}^d \to \mathbb{R}^k$ and $\psi : \mathbb{R}^d \to \mathbb{R}^l$ for some $k, l \in \mathbb{N}$, $d^{\text{UKP}}_{g_\lambda,K,\mathcal{L}^p}(\phi,\psi) \geq 0$,*
3. *(Symmetric) For any two functions $\phi : \mathbb{R}^d \to \mathbb{R}^k$ and $\psi : \mathbb{R}^d \to \mathbb{R}^l$ for some $k, l \in \mathbb{N}$, $d^{\text{UKP}}_{g_\lambda,K,\mathcal{L}^p}(\phi,\psi) = d^{\text{UKP}}_{g_\lambda,K,\mathcal{L}^p}(\psi,\phi)$,*
4. *(Triangle inequality) For any three functions $\phi : \mathbb{R}^d \to \mathbb{R}^k$, $\psi : \mathbb{R}^d \to \mathbb{R}^l$ and $\varphi : \mathbb{R}^d \to \mathbb{R}^m$ for some $k, l, m \in \mathbb{N}$, $d^{\text{UKP}}_{g_\lambda,K,\mathcal{L}^p}(\phi,\psi) \leq d^{\text{UKP}}_{g_\lambda,K,\mathcal{L}^p}(\phi,\varphi) + d^{\text{UKP}}_{g_\lambda,K,\mathcal{L}^p}(\varphi,\psi)$.*

Hence, $d^{\text{UKP}}_{g_\lambda,K,\mathcal{L}^p}(\phi,\psi)$ is a pseudometric over the space of all functions that maps $\mathbb{R}^d$ to some Euclidean space $\mathbb{R}^t$ for any $t \in \mathbb{N}$.

Proof. Equation 13 and the 4 pseudometric properties readily follow from the expression of $d^{UKP}_{g_\lambda,K,\mathcal{L}^p}(\phi,\psi)$ in Definition 4 in terms of the p-Schatten norm of $[g_\lambda(\mathcal{T}_\phi)\mathcal{T}_\phi - g_\lambda(\mathcal{T}_\psi)\mathcal{T}_\psi]$ and using the properties of p-Schatten norms as given in Proposition 2.1 of Pfeiffer (2021). □

We now analyze the invariance properties of the pseudometric $d^{UKP}_{g_\lambda,K,\mathcal{L}^p}$ and identify the transformations of the representations ϕ and ψ that leave its value unchanged. Based on the following theorem, we can identify representations that UKP treats as equivalent in terms of prediction-based performance for a general collection of kernel ridge regression tasks corresponding to a particular kernel K.

Remark 4. *One of the most novel aspect of our contributions is the exact identification of the mathematical relationship between representations that lead them to have the same generalization performance. To be specific, the GULP paper derives in their Theorem 1 and Theorem 2 that two representations and are equivalent from the lens of the GULP metric if and only if one can be expressed as an orthogonal linear transformation of the other. However, the authors of the GULP paper do not discuss why orthogonality of representations plays such a crucial role in their results. Our results clearly show that the source of the appearance of the orthogonality condition is the choice of the similarity function in the GULP paper, which is the linear kernel . We are able to delineate the exact relationship between the choice of invariance and the choice of the kernel, thus extending their results to a much broader domain. Further, the proof techniques used to prove the invariance properties of the GULP metric in Theorem 2 are specific to the case of the linear kernel and rely on complicated manipulations based on linear algebra theory (such as analyzing homogeneous Sylvester equations). In contrast, our proofs for the invariance properties of the UKP metric use standard and systematic functional analysis techniques in RKHSs. Therefore, the proofs are easier to understand and generalize the results available in the GULP paper. Moreover, we emphasize that no closed-form expression of the population version of the pseudometric was given in the GULP paper. The authors of the GULP paper only provided the closed-form expression of the estimator of the pseudometric, and that too only for the linear kernel. Therefore, the preliminary KRR discussion in the GULP paper considers only a specific case of the plug-in estimator we propose in our paper, and express in a more computationally tractable form in terms of Gram matrices.*

Theorem 9. *Assume that the setting of Theorem 7 holds true. Then, for any $p \geq 1$ and $\lambda > 0$, given any two representations $\phi : \mathbb{R}^d \to \mathbb{R}^k$ and $\psi : \mathbb{R}^d \to \mathbb{R}^l$ we have that*

$$d^{\text{UKP}}_{g_\lambda, K, \mathcal{L}^p}(\phi, \psi) = 0 \quad \textit{if and only if} \quad \mathcal{T}_\phi = \mathcal{T}_\psi. \tag{14}$$

Further, let $\mathcal{H}$ be the class of transformations under which the kernel K is invariant, i.e., $\mathcal{H} = \{h : K(\cdot, \cdot) = K(h(\cdot), h(\cdot)) \text{ a.e. } P_X\}$. Then, the UKP *distance $d^{\text{UKP}}_{g_\lambda, K, \mathcal{L}^p}(\phi, \psi)$ between representations $\phi(X)$ and $\psi(X)$ is invariant under the same class of transformations that the kernel K is invariant for, i.e., for any $h_1, h_2 \in \mathcal{H}$,*

$$d^{\text{UKP}}_{g_\lambda, K, \mathcal{L}^p}(h_1 \circ \phi, h_2 \circ \psi) = d^{\text{UKP}}_{g_\lambda, K, \mathcal{L}^p}(\phi, \psi)$$

and if either h_1 or h_2 does not belong to $\mathcal{H}$,

$$d^{\text{UKP}}_{g_\lambda, K, \mathcal{L}^p}(h_1 \circ \phi, h_2 \circ \psi) \neq d^{\text{UKP}}_{g_\lambda, K, \mathcal{L}^p}(\phi, \psi).$$

Consequently, a necessary and sufficient condition for the UKP *distance $d^{\text{UKP}}_{g_\lambda, K, \mathcal{L}^p}(\phi, \psi)$ between representations $\phi(X)$ and $\psi(X)$ to be zero is that $K_\phi(\cdot, \cdot) = K_\psi(\cdot, \cdot)$ a.e. P_X.*

Proof. The sufficiency part of the claim in Equation 14 is trivial. For the necessity part, we observe that, by the definiteness of p-Schatten norms (follows readily from the definiteness of 1-Schatten norm i.e trace norm, see Lemma 8 of Chapter 3 of Schatten (2013)), we must have that

$$g_\lambda(\mathcal{T}_\phi)\mathcal{T}_\phi = g_\lambda(\mathcal{T}_\psi)\mathcal{T}_\psi. \tag{15}$$

Under the given conditions on the kernel K, the integral operators $\mathcal{T}_\phi$ and $\mathcal{T}_\psi$ corresponding to the kernels K_ϕ and K_ψ both admit spectral decompositions. Let $\left(\mu_i^\phi, e_i^\phi\right)_{i=1}^\infty$ and $\left(\mu_j^\psi, e_j^\psi\right)_{j=1}^\infty$ be the eigenvalue-eigenfunction pairs corresponding to the spectral decomposition of $\mathcal{T}_\phi$ and $\mathcal{T}_\psi$, respectively. Then, we have that

$$\mathcal{T}_\phi = \sum_{i=1}^\infty \mu_i^\phi \left(e_i^\phi \otimes_{L^2(P_X)} e_i^\phi\right)$$

and

$$\mathcal{T}_\psi = \sum_{j=1}^\infty \mu_j^\psi \left(e_j^\psi \otimes_{L^2(P_X)} e_j^\psi\right).$$

Since K is a positive definite, symmetric, continuous and bounded kernel defined on a separable domain, $\mathcal{T}_\phi$ and $\mathcal{T}_\psi$ are compact, self-adjoint, trace-class operators. Therefore, we must have that $\mu_i^\phi, \mu_j^\psi > 0$ and $\lim_{i\to\infty} \mu_i^\phi = \lim_{j\to\infty} \mu_j^\psi = 0$. Further, $(e_i^\phi)_{i=1}^\infty$ and $(e_j^\psi)_{j=1}^\infty$ constitute a pair of orthonormal bases of $\text{Ran}(\mathcal{T}_\phi)$ and $\text{Ran}(\mathcal{T}_\psi)$, respectively.

Consequently, it can be readily verified that

$$g_\lambda(\mathcal{T}_\phi)\mathcal{T}_\phi = \sum_{i=1}^\infty \mu_i^\phi g_\lambda(\mu_i^\phi) \left(e_i^\phi \otimes_{L^2(P_X)} e_i^\phi\right)$$

and

$$g_\lambda(\mathcal{T}_\psi)\mathcal{T}_\psi = \sum_{j=1}^\infty \mu_j^\psi g_\lambda(\mu_j^\psi) \left(e_j^\psi \otimes_{L^2(P_X)} e_j^\psi\right).$$

Therefore, we must have that

$$\begin{aligned} & g_\lambda(\mathcal{T}_\phi)\mathcal{T}_\phi = g_\lambda(\mathcal{T}_\psi)\mathcal{T}_\psi \\ \iff & \sum_{i=1}^\infty \mu_i^\phi g_\lambda(\mu_i^\phi) \left(e_i^\phi \otimes_{L^2(P_X)} e_i^\phi\right) = \sum_{j=1}^\infty \mu_j^\psi g_\lambda(\mu_j^\psi) \left(e_j^\psi \otimes_{L^2(P_X)} e_j^\psi\right). \end{aligned} \tag{16}$$

Define $t_{ij} \coloneqq \left\langle e_i^{\phi}, e_j^{\psi} \right\rangle_{L^2(P_X)}$ for all i, j. Further, define $V_i = \{j \in \mathbb{N} : t_{ij} \neq 0\}$ for all i and $W_j = \{i \in \mathbb{N} : t_{ij} \neq 0\}$ for all j.

Now, taking the $L^2(P_X)$ inner product of both the RHS and LHS of equation 16 with e_j^{ψ}, we have that

$$\sum_{i=1}^{\infty} \mu_i^{\phi} g_\lambda(\mu_i^{\phi}) t_{ij} e_i^{\phi}(\cdot) = \mu_j^{\psi} g_\lambda(\mu_j^{\psi}) e_j^{\psi}(\cdot). \tag{17}$$

Taking the $L^2(P_X)$ inner product of both the RHS and LHS of equation 17 with e_k^{ϕ}, we have that

$$\begin{aligned} &\mu_k^{\phi} g_\lambda(\mu_k^{\phi}) t_{kj} = \mu_j^{\psi} g_\lambda(\mu_j^{\psi}) t_{kj} \\ \iff &t_{kj} \left(\mu_k^{\phi} g_\lambda(\mu_k^{\phi}) - \mu_j^{\psi} g_\lambda(\mu_j^{\psi}) \right) = 0. \end{aligned} \tag{18}$$

Taking the $L^2(P_X)$ inner product of both the RHS and LHS of equation 17 with e_k^{ψ}, we have that

$$\begin{aligned} &\sum_{i=1}^{\infty} \mu_i^{\phi} g_\lambda(\mu_i^{\phi}) t_{ij} t_{ik} = \begin{cases} \mu_j^{\psi} g_\lambda(\mu_j^{\psi}) \text{ if } j = k \\ 0, \text{ if } j \neq k \end{cases} \\ \iff &\sum_{i \in W_j \cap W_k} \mu_i^{\phi} g_\lambda(\mu_i^{\phi}) t_{ij} t_{ik} = \begin{cases} \mu_j^{\psi} g_\lambda(\mu_j^{\psi}) \text{ if } j = k \\ 0, \text{ if } j \neq k \end{cases} . \end{aligned} \tag{19}$$

Using equation 18 and equation 19, we have that

$$\mu_j^{\psi} g_\lambda(\mu_j^{\psi}) \left[\sum_{i \in W_j} t_{ij}^2 - 1 \right] = 0 \tag{20}$$

and, if $j \neq k$,

$$\mu_j^{\psi} g_\lambda(\mu_j^{\psi}) \left(\sum_{i \in W_j \cap W_k} t_{ij} t_{ik} \right) = 0. \tag{21}$$

Therefore, from equation 20 and equation 21, we obtain that

$$\sum_{i \in W_j} t_{ij}^2 = 1 \tag{22}$$

and, if $j \neq k$,

$$\sum_{i \in W_j \cap W_k} t_{ij} t_{ik} = 0. \tag{23}$$

In exactly analogous manner, we can also obtain

$$\sum_{j \in V_i} t_{ij}^2 = 1 \tag{24}$$

and, if $i \neq k$,

$$\sum_{j \in V_i \cap V_k} t_{ij} t_{kj} = 0. \tag{25}$$

Note that $(e_j^{\psi})_{j=1}^{\infty}$ can be extended to obtain an orthonormal basis for $L^2(P_X)$. Let $B = \left\{ \cup_{j=1}^{\infty} e_j^{\psi} \right\} \cup \left\{ \cup_{l=1}^{\infty} z_l^{\psi} \right\}$ be the resulting orthonormal basis of $L^2(P_X)$ obtained by said extension. Now,

$$e_i^{\phi} = \sum_{j=1}^{\infty} \left\langle e_i^{\phi}, e_j^{\psi} \right\rangle e_j^{\psi} + \sum_{l=1}^{\infty} \left\langle e_i^{\phi}, z_l^{\psi} \right\rangle z_l^{\psi}. \tag{26}$$

Therefore, using equation 26 and equation 24 along with the orthonormality of $(e_i^\phi)_{i=1}^\infty$, we have,

$$\begin{aligned}
&\left\|e_i^\phi\right\|_{L^2(P_X)} = 1 \\
\iff &\sum_{j=1}^{\infty} \left\langle e_i^\phi, e_j^\psi \right\rangle^2 + \sum_{l=1}^{\infty} \left\langle e_i^\phi, z_l^\psi \right\rangle^2 = 1 \\
\iff &\sum_{j\in V_i} t_{ij}^2 + \sum_{l=1}^{\infty} \left\langle e_i^\phi, z_l^\psi \right\rangle^2 = 1 \\
\iff &\sum_{l=1}^{\infty} \left\langle e_i^\phi, z_l^\psi \right\rangle^2 = 0 \\
\iff &\left\langle e_i^\phi, z_l^\psi \right\rangle = 0 \text{ for all l and i.}
\end{aligned}$$

Hence, for all i, $e_i^\phi \in \text{Span}\left\{e_j^\psi, j \in \mathbb{N}\right\}$. Consequently, $\mathcal{T}_\phi g_\lambda(\mathcal{T}_\phi) e_j^\psi = \sum_{i=1}^\infty \mu_i^\phi g_\lambda(\mu_i^\phi) t_{ij} e_i^\phi \in \text{Span}\left\{e_i^\phi, i \in \mathbb{N}\right\} \subset \text{Span}\left\{e_j^\psi, j \in \mathbb{N}\right\}$.

Now, using equation 23 and equation 18, for any $j \neq k$, we have

$$\begin{aligned}
\left\langle \mathcal{T}_\phi g_\lambda(\mathcal{T}_\phi) e_j^\psi, e_k^\psi \right\rangle_{L^2(P_X)} =& \sum_{i=1}^{\infty} \mu_i^\phi g_\lambda(\mu_i^\phi) t_{ij} t_{ik} \\
=& \sum_{i\in W_j \cap W_k} \mu_i^\phi g_\lambda(\mu_i^\phi) t_{ij} t_{ik} \\
=& \mu_j^\psi g_\lambda(\mu_j^\psi) \sum_{i\in W_j\cap W_k} t_{ij} t_{ik} \\
=& 0.
\end{aligned}$$

Finally, using equation 20 and equation 18, we have that

$$\begin{aligned}
\left\langle \mathcal{T}_\phi g_\lambda(\mathcal{T}_\phi) e_j^\psi, e_j^\psi \right\rangle_{L^2(P_X)} =& \sum_{i=1}^{\infty} \mu_i^\psi g_\lambda(\mu_i^\psi) t_{ij}^2 \\
=& \sum_{i\in W_j} \mu_i^\psi g_\lambda(\mu_i^\psi) t_{ij}^2 \\
=& \mu_j^\psi g_\lambda(\mu_j^\psi) \sum_{i\in W_j} t_{ij}^2 \\
=& \mu_j^\psi g_\lambda(\mu_j^\psi) > 0.
\end{aligned}$$

Therefore, $\mathcal{T}_\phi g_\lambda(\mathcal{T}_\phi) e_j^\phi = \mu_j^\psi g_\lambda(\mu_j^\psi) e_j^\psi$ for all j. Therefore, all the eigenfunctions of $\mathcal{T}_\psi g_\lambda(\mathcal{T}_\psi)$ (and hence those of $\mathcal{T}_\psi$) are also eigenfunctions of $\mathcal{T}_\phi g_\lambda(\mathcal{T}_\phi)$ (and hence those of $\mathcal{T}_\phi$). By symmetry, all the eigenfunctions of $\mathcal{T}_\phi$ are also eigenfunctions of $\mathcal{T}_\psi$. Therefore, $\mathcal{T}_\phi$ and $\mathcal{T}_\psi$ (equivalently, $\mathcal{T}_\phi g_\lambda(\mathcal{T}_\phi)$ and $\mathcal{T}_\psi g_\lambda(\mathcal{T}_\psi)$) have exactly the same eigenfunctions.

Consequently, equation 16 can be now written as

$$\begin{aligned}
&g_\lambda(\mathcal{T}_\phi)\mathcal{T}_\phi = g_\lambda(\mathcal{T}_\psi)\mathcal{T}_\psi \\
\iff &\sum_{i=1}^{\infty} \mu_i^\phi g_\lambda(\mu_i^\phi) \left(e_i^\phi \otimes_{L^2(P_X)} e_i^\phi\right) = \sum_{i=1}^{\infty} \mu_i^\psi g_\lambda(\mu_i^\psi) \left(e_i^\phi \otimes_{L^2(P_X)} e_i^\phi\right). \qquad (27)
\end{aligned}$$

Taking the $L^2(P_X)$ inner product of both the RHS and LHS of equation 27 with e_i^ϕ twice and using the injectivity of $x \mapsto x g_\lambda(x)$ over $x \in [0, \kappa]$, we have that, for any i,

$$\begin{aligned}
&\mu_i^\phi g_\lambda(\mu_i^\phi) = \mu_i^\psi g_\lambda(\mu_i^\psi) \\
\iff &\mu_i^\phi = \mu_i^\psi.
\end{aligned}$$

Therefore, we must have that the integral operators $\mathcal{T}_\phi$ and $\mathcal{T}_\psi$ have the same spectral decomposition. Consequently, we must have that

$$\mathcal{T}_\phi = \mathcal{T}_\psi. \tag{28}$$

This concludes the proof of the necessity part.

Equation 28 is equivalent to the following condition : For any $f \in L^2(P_X)$,

$$\int \left[K_\phi(\cdot, x) f(x) dP_X(x) - K_\psi(\cdot, x)\right] f(x) dP_X(x) = 0. \tag{29}$$

Consequently, Equation 28 is equivalent to the condition $K_\phi(\cdot,\cdot) = K_\psi(\cdot,\cdot)$ a.e. P_X.

On the other hand, if K_ϕ and K_ψ differ on a set of positive measure under P_X, then clearly $\mathcal{T}_\phi$ and $\mathcal{T}_\psi$ define two distinct operators. Using the injectivity of $x \mapsto x g_\lambda(x)$, this ensures that $\mathcal{T}_\phi g_\lambda(\mathcal{T}_\phi) \neq \mathcal{T}_\psi g_\lambda(\mathcal{T}_\psi)$, leading to $d^{UKP}_{g_\lambda, K, \mathcal{L}^p}(\phi, \psi) > 0$. This completes the proof of the required result. □

A.6 Properties of $d^{UKP}_{g_\lambda, K, \mathcal{L}^2}$

Theorem 10. *Assume that the setting of Theorem 7 holds true. Then, for any $\lambda > 0$, the squared* UKP *distance $d^{\text{UKP}}_{g_\lambda, K, \mathcal{L}^2}(\phi, \psi)$ between representations $\phi(X)$ and $\psi(X)$ can be expressed as*

$$\left[d^{\text{UKP}}_{g_\lambda, K, \mathcal{L}^2}(\phi, \psi)\right]^2 = \operatorname{Tr}\left(g_\lambda(\Sigma_\phi)\Sigma_\phi g_\lambda(\Sigma_\phi)\Sigma_\phi\right) + \operatorname{Tr}\left(g_\lambda(\Sigma_\psi)\Sigma_\psi g_\lambda(\Sigma_\psi)\Sigma_\psi\right) - 2\operatorname{Tr}\left(g_\lambda(\Sigma_\phi)\Sigma_{\phi\psi} g_\lambda(\Sigma_\psi)\Sigma_{\psi\phi}\right).$$

In particular, when the chosen spectral regularizer is the Tikhonov regularizer $g^{\text{Tik}}_\lambda(x) = \frac{1}{x+\lambda}$, the squared UKP distance $d^{\text{UKP}}_{\lambda, K, \mathcal{L}^2}(\phi, \psi)$ can be expressed as

$$\left[d^{\text{UKP}}_{\lambda, K, \mathcal{L}^2}(\phi, \psi)\right]^2 = \operatorname{Tr}\left(\Sigma_\phi^{-\lambda}\Sigma_\phi\Sigma_\phi^{-\lambda}\Sigma_\phi\right) + \operatorname{Tr}\left(\Sigma_\psi^{-\lambda}\Sigma_\psi\Sigma_\psi^{-\lambda}\Sigma_\psi\right) - 2\operatorname{Tr}\left(\Sigma_\phi^{-\lambda}\Sigma_{\phi\psi}\Sigma_\psi^{-\lambda}\Sigma_{\psi\phi}\right).$$

Proof. Using Definition 4 and the results of Lemma 1, we have that the squared UKP distance $d^{UKP}_{g_\lambda, K, \mathcal{L}^2}(\phi, \psi)$ between representations $\phi(X)$ and $\psi(X)$ can be expressed as

$$\begin{aligned}
\left[d^{UKP}_{g_\lambda, K, \mathcal{L}^2}(\phi, \psi)\right]^2 &= \|g_\lambda(\mathcal{T}_\phi)\mathcal{T}_\phi - g_\lambda(\mathcal{T}_\psi)\mathcal{T}_\psi\|^2_{\mathcal{L}^2(L^2(P_X))} \\
&= \|\mathcal{T}_\phi g_\lambda(\mathcal{T}_\phi) - \mathcal{T}_\psi g_\lambda(\mathcal{T}_\psi)\|^2_{\mathcal{L}^2(L^2(P_X))} \\
&= \left\|\mathfrak{I}_\phi g_\lambda(\Sigma_\phi)\mathfrak{I}_\phi^* - \mathfrak{I}_\psi g_\lambda(\Sigma_\psi)\mathfrak{I}_\psi^*\right\|^2_{\mathcal{L}^2(L^2(P_X))} \\
&= \left\langle \mathfrak{I}_\phi g_\lambda(\Sigma_\phi)\mathfrak{I}_\phi^*, \mathfrak{I}_\phi g_\lambda(\Sigma_\phi)\mathfrak{I}_\phi^* \right\rangle_{\mathcal{L}^2(L^2(P_X))} + \left\langle \mathfrak{I}_\psi g_\lambda(\Sigma_\psi)\mathfrak{I}_\psi^*, \mathfrak{I}_\psi g_\lambda(\Sigma_\psi)\mathfrak{I}_\psi^* \right\rangle_{\mathcal{L}^2(L^2(P_X))} \\
&\quad - 2\left\langle \mathfrak{I}_\phi g_\lambda(\Sigma_\phi)\mathfrak{I}_\phi^*, \mathfrak{I}_\psi g_\lambda(\Sigma_\psi)\mathfrak{I}_\psi^* \right\rangle_{\mathcal{L}^2(L^2(P_X))} \\
&= \operatorname{Tr}\left(g_\lambda(\Sigma_\phi)\Sigma_\phi g_\lambda(\Sigma_\phi)\Sigma_\phi\right) + \operatorname{Tr}\left(g_\lambda(\Sigma_\psi)\Sigma_\psi g_\lambda(\Sigma_\psi)\Sigma_\psi\right) - 2\operatorname{Tr}\left(g_\lambda(\Sigma_\phi)\Sigma_{\phi\psi} g_\lambda(\Sigma_\psi)\Sigma_{\psi\phi}\right).
\end{aligned}$$

When the chosen spectral regularizer is the Tikhonov regularizer $g^{\text{Tik}}_\lambda(x) = \frac{1}{x+\lambda}$, the UKP distance $d^{UKP}_{\lambda, K, \mathcal{L}^2}(\phi, \psi) \coloneqq d^{UKP}_{g^{\text{Tik}}_\lambda, K, \mathcal{L}^2}(\phi, \psi)$ can be derived from the above computation. □

A.7 Estimation of $d^{UKP}_{g_\lambda, K, \mathcal{L}^2}$ and finite-sample concentration of the estimator

Using the plugin-estimators $\hat{\Sigma}_\phi$, $\hat{\Sigma}_\psi$, $\hat{\Sigma}_{\phi\psi}$ and $\hat{\Sigma}_{\psi\phi}$ for estimating their corresponding population counterparts, we can obtain a V-statistic type estimator for $d^{UKP}_{g_\lambda, K, \mathcal{L}^2}$ based on i.i.d samples $X_1, \ldots, X_n$ drawn from P_X, and is given by

$$\begin{aligned}
&\hat{d}^{UKP}_{g_\lambda, K, \mathcal{L}^2}(\phi, \psi) \\
&= \left[\operatorname{Tr}\left(g_\lambda(\Sigma_\phi)\Sigma_\phi g_\lambda(\Sigma_\phi)\Sigma_\phi\right) + \operatorname{Tr}\left(g_\lambda(\Sigma_\psi)\Sigma_\psi g_\lambda(\Sigma_\psi)\Sigma_\psi\right) - 2\operatorname{Tr}\left(g_\lambda(\Sigma_\phi)\Sigma_{\phi\psi} g_\lambda(\Sigma_\psi)\Sigma_{\psi\phi}\right)\right]^{\frac{1}{2}}.
\end{aligned} \tag{30}$$

Using the results in Lemma 2, one can prove an equivalent expression of the estimator $\hat{d}^{UKP}_{g_\lambda, K, \mathcal{L}^2}(\phi, \psi)$ in terms of the kernel Gram matrices $K_{n,\phi}$ and $K_{n,\psi}$, which is more useful in practice and implementation.

Theorem 11. *For any $\lambda > 0$, the V-statistic type estimator $\hat{d}^{\text{UKP}}_{g_\lambda,K,\mathcal{L}^2}(\phi,\psi)$ of $d^{\text{UKP}}_{g_\lambda,K,\mathcal{L}^2}(\phi,\psi)$ between representations $\phi(X)$ and $\psi(X)$ can be expressed as*

$$\begin{aligned}
&\hat{d}^{\text{UKP}}_{g_\lambda,K,\mathcal{L}^2}(\phi,\psi)\\
=&\frac{1}{n}\left[\text{Tr}\left(K_{n,\phi}g_\lambda(\frac{1}{n}K_{n,\phi}+\lambda I)K_{n,\phi}g_\lambda(\frac{1}{n}K_{n,\phi}+\lambda I)\right)+\text{Tr}\left(K_{n,\psi}g_\lambda(\frac{1}{n}K_{n,\psi}+\lambda I)K_{n,\psi}g_\lambda(\frac{1}{n}K_{n,\psi}+\lambda I)\right)\right.\\
&\left.-2\ \text{Tr}\left(K_{n,\phi}g_\lambda(\frac{1}{n}K_{n,\phi}K_{n,\psi}g_\lambda(\frac{1}{n}K_{n,\psi})\right)\right]^{\frac{1}{2}}.
\end{aligned}$$

When the chosen spectral regularizer is the Tikhonov regularizer $g^{\text{Tik}}_\lambda(x)=\frac{1}{x+\lambda}$, the estimator $\hat{d}^{\text{UKP}}_{\lambda,K,\mathcal{L}^2}(\phi,\psi)$ of $d^{\text{UKP}}_{\lambda,K,\mathcal{L}^2}(\phi,\psi)$ is given by

$$\begin{aligned}
&\hat{d}^{\text{UKP}}_{\lambda,K,\mathcal{L}^2}(\phi,\psi)\\
=&\left[\text{Tr}\left(K_{n,\phi}(K_{n,\phi}+n\lambda I)^{-1}K_{n,\phi}(K_{n,\phi}+n\lambda I)^{-1}\right)+\text{Tr}\left(K_{n,\psi}(K_{n,\psi}+n\lambda I)^{-1}K_{n,\psi}(K_{n,\psi}+n\lambda I)^{-1}\right)\right.\\
&\left.-2\ \text{Tr}\left(K_{n,\phi}(K_{n,\phi}+n\lambda I)^{-1}K_{n,\psi}(K_{n,\psi}+n\lambda I)^{-1}\right)\right]^{\frac{1}{2}}.
\end{aligned}$$

We are also able to prove a finite-sample concentration inequality for the estimator $\hat{d}^{UKP}_{\lambda,K,\mathcal{L}^2}(\phi,\psi)$ for $\lambda>0$ that demonstrates that the statistical estimation error of $d^{UKP}_{\lambda,K,\mathcal{L}^2}(\phi,\psi)$ converges to 0 at the parametric rate $n^{-\frac{1}{2}}$, where n is the sample size. For simplicity, the concentration inequality is derived for the specific regularizer that corresponds to kernel ridge regression (Tikhonov regularization), but similar rates of convergence (in terms of sample size n) will hold for other choices of the regularizer g_λ with possible different dependence on λ.

Proof of Theorem 6

Proof. Note that for any $x,y\in\mathbb{R}^d$, $\hat{\Sigma}^{-\lambda}_\phi\left[K_\phi(\cdot,x)\otimes_{\mathcal{H}_\phi}K_\phi(\cdot,x)\right]\hat{\Sigma}^{-\lambda}_\phi\left[K_\phi(\cdot,y)\otimes_{\mathcal{H}_\phi}K_\phi(\cdot,y)\right]$ is a rank-one operator with eigenvalue $\left\langle\hat{\Sigma}^{-\frac{\lambda}{2}}_\phi K_\phi(\cdot,x),\hat{\Sigma}^{-\frac{\lambda}{2}}_\phi K_\phi(\cdot,y)\right\rangle^2_{\mathcal{H}_\phi}$ and eigenfunction $\frac{\hat{\Sigma}^{-\lambda}_\phi K_\phi(\cdot,x)}{\left\|\hat{\Sigma}^{-\lambda}_\phi K_\phi(\cdot,x)\right\|_{\mathcal{H}_\phi}}$. Similarly, $\hat{\Sigma}^{-\lambda}_\psi\left[K_\psi(\cdot,x)\otimes_{\mathcal{H}_\psi}K_\psi(\cdot,x)\right]\hat{\Sigma}^{-\lambda}_\psi\left[K_\psi(\cdot,y)\otimes_{\mathcal{H}_\psi}K_\psi(\cdot,y)\right]$ is a rank-one operator with eigenvalue $\left\langle\hat{\Sigma}^{-\frac{\lambda}{2}}_\psi K_\psi(\cdot,x),\hat{\Sigma}^{-\frac{\lambda}{2}}_\psi K_\psi(\cdot,y)\right\rangle^2_{\mathcal{H}_\psi}$ and eigenfunction $\frac{\hat{\Sigma}^{-\lambda}_\psi K_\psi(\cdot,x)}{\left\|\hat{\Sigma}^{-\lambda}_\psi K_\psi(\cdot,x)\right\|_{\mathcal{H}_\psi}}$. Further,
$\hat{\Sigma}^{-\lambda}_\phi\left[K_\phi(\cdot,x)\otimes_{\mathcal{L}^2(\mathcal{H}_\phi,\mathcal{H}_\psi)}K_\psi(\cdot,x)\right]\times\hat{\Sigma}^{-\lambda}_\psi\left[K_\psi(\cdot,y)\otimes_{\mathcal{L}^2(\mathcal{H}_\phi,\mathcal{H}_\psi)}K_\phi(\cdot,y)\right]$ is a rank-one operator with eigenvalue $\left\langle\hat{\Sigma}^{-\frac{\lambda}{2}}_\phi K_\phi(\cdot,x),\hat{\Sigma}^{-\frac{\lambda}{2}}_\phi K_\phi(\cdot,y)\right\rangle_{\mathcal{H}_\phi}\times\left\langle\hat{\Sigma}^{-\frac{\lambda}{2}}_\psi K_\psi(\cdot,x),\hat{\Sigma}^{-\frac{\lambda}{2}}_\psi K_\psi(\cdot,y)\right\rangle_{\mathcal{H}_\psi}$ and eigenfunction $\frac{\hat{\Sigma}^{-\lambda}_\phi K_\phi(\cdot,x)}{\left\|\hat{\Sigma}^{-\lambda}_\phi K_\phi(\cdot,x)\right\|_{\mathcal{H}_\phi}}$.

Using these facts, we have that the squared V-statistic type estimator of $d^{\text{UKP}}_{\lambda,K}$ can be expressed as

$$\begin{aligned}
&\left[\hat{d}^{\text{UKP}}_\lambda(\phi,\psi)\right]^2\\
=&\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n\left[\left\langle\hat{\Sigma}^{-\frac{\lambda}{2}}_\phi K_\phi(\cdot,X_i),\hat{\Sigma}^{-\frac{\lambda}{2}}_\phi K_\phi(\cdot,X_j)\right\rangle_{\mathcal{H}_\phi}-\left\langle\hat{\Sigma}^{-\frac{\lambda}{2}}_\psi K_\psi(\cdot,X_i),\hat{\Sigma}^{-\frac{\lambda}{2}}_\psi K_\psi(\cdot,X_j)\right\rangle_{\mathcal{H}_\psi}\right]^2.
\end{aligned}$$

Let us define the following quantity

$$\begin{aligned}
&\left[\tilde{d}^{\text{UKP}}_\lambda(\phi,\psi)\right]^2\\
:=&\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n\left[\left\langle\Sigma^{-\frac{\lambda}{2}}_\phi K_\phi(\cdot,X_i),\Sigma^{-\frac{\lambda}{2}}_\phi K_\phi(\cdot,X_j)\right\rangle_{\mathcal{H}_\phi}-\left\langle\Sigma^{-\frac{\lambda}{2}}_\psi K_\psi(\cdot,X_i),\Sigma^{-\frac{\lambda}{2}}_\psi K_\psi(\cdot,X_j)\right\rangle_{\mathcal{H}_\psi}\right]^2
\end{aligned}$$

which is $\left[\hat{d}^{\text{UKP}}_{\lambda}(\phi,\psi)\right]^2$ with $\hat{\Sigma}_\phi$ and $\hat{\Sigma}_\phi$ replaced by Σ_ϕ and Σ_ϕ, respectively. We utilize the triangle inequality to bound the difference between the squared V-statistic type estimator $\left[\hat{d}^{\text{UKP}}_{\lambda}(\phi,\psi)\right]^2$ and the squared population distance $\left[d^{\text{UKP}}_{\lambda,K}(\phi,\psi)\right]^2$ as follows:

$$\begin{aligned}&\left|\left[\hat{d}^{\text{UKP}}_{\lambda}(\phi,\psi)\right]^2-\left[d^{\text{UKP}}_{\lambda,K}(\phi,\psi)\right]^2\right|\\&\leq\underbrace{\left|\left[\hat{d}^{\text{UKP}}_{\lambda}(\phi,\psi)\right]^2-\left[\tilde{d}^{\text{UKP}}_{\lambda}(\phi,\psi)\right]^2\right|}_{\mathbf{A}}+\underbrace{\left|\left[\tilde{d}^{\text{UKP}}_{\lambda}(\phi,\psi)\right]^2-\left[d^{\text{UKP}}_{\lambda,K}(\phi,\psi)\right]^2\right|}_{\mathbf{B}}.\end{aligned}\tag{31}$$

We now proceed to bound $\mathbf{A}$. Let us define

$$\begin{aligned}\hat{A}_{ij,\phi}&=\left\langle\hat{\Sigma}_\phi^{-\frac{\lambda}{2}}K_\phi(\cdot,X_i),\hat{\Sigma}_\phi^{-\frac{\lambda}{2}}K_\phi(\cdot,X_j)\right\rangle_{\mathcal{H}_\phi},\\A_{ij,\phi}&=\left\langle\Sigma_\phi^{-\frac{\lambda}{2}}K_\phi(\cdot,X_i),\Sigma_\phi^{-\frac{\lambda}{2}}K_\phi(\cdot,X_j)\right\rangle_{\mathcal{H}_\phi},\\\hat{A}_{ij,\psi}&=\left\langle\hat{\Sigma}_\psi^{-\frac{\lambda}{2}}K_\psi(\cdot,X_i),\hat{\Sigma}_\psi^{-\frac{\lambda}{2}}K_\psi(\cdot,X_j)\right\rangle_{\mathcal{H}_\psi},\\A_{ij,\psi}&=\left\langle\Sigma_\psi^{-\frac{\lambda}{2}}K_\psi(\cdot,X_i),\Sigma_\psi^{-\frac{\lambda}{2}}K_\psi(\cdot,X_j)\right\rangle_{\mathcal{H}_\psi}.\end{aligned}$$

Then, we have that

$$\left|\hat{A}_{ij,\phi}\right|\leq\|K_\phi(\cdot,X_i)\|^2_{\mathcal{H}_\phi}\times\left\|\hat{\Sigma}_\phi^{-\frac{\lambda}{2}}\right\|^2_{\mathcal{L}^\infty(\mathcal{H}_\phi)}\leq\frac{\kappa}{\lambda}.$$

Similarly, we can show that $\left|\hat{A}_{ij,\psi}\right|\leq\frac{\kappa}{\lambda}$,$|A_{ij,\phi}|\leq\frac{\kappa}{\lambda}$ and $|A_{ij,\psi}|\leq\frac{\kappa}{\lambda}$. Now, we have that

$$\begin{aligned}\left|\hat{A}_{ij,\phi}-A_{ij,\phi}\right|&=\left|\left\langle K_\phi(\cdot,X_i),\left(\hat{\Sigma}_\phi^{-\lambda}-\Sigma_\phi^{-\lambda}\right)K_\phi(\cdot,X_j)\right\rangle_{\mathcal{H}_\phi}\right|\\&\leq\kappa\left\|\hat{\Sigma}_\phi^{-\lambda}-\Sigma_\phi^{-\lambda}\right\|_{\mathcal{L}^\infty(\mathcal{H}_\phi)}\leq\kappa\left\|\hat{\Sigma}_\phi^{-\lambda}-\Sigma_\phi^{-\lambda}\right\|_{\mathcal{L}^2(\mathcal{H}_\phi)}.\end{aligned}$$

Similarly, we have that

$$\begin{aligned}\left|\hat{A}_{ij,\psi}-A_{ij,\psi}\right|&=\left|\left\langle K_\psi(\cdot,X_i),\left(\hat{\Sigma}_\psi^{-\lambda}-\Sigma_\psi^{-\lambda}\right)K_\psi(\cdot,X_j)\right\rangle_{\mathcal{H}_\psi}\right|\\&\leq\kappa\left\|\hat{\Sigma}_\psi^{-\lambda}-\Sigma_\psi^{-\lambda}\right\|_{\mathcal{L}^\infty(\mathcal{H}_\psi)}\leq\kappa\left\|\hat{\Sigma}_\psi^{-\lambda}-\Sigma_\psi^{-\lambda}\right\|_{\mathcal{L}^2(\mathcal{H}_\psi)}.\end{aligned}$$

Note that,

$$\begin{aligned}&\left\|\hat{\Sigma}_\phi^{-\lambda}-\Sigma_\phi^{-\lambda}\right\|_{\mathcal{L}^\infty(\mathcal{H}_\phi)}\\&=\left\|\left(\hat{\Sigma}_\phi+\lambda I\right)^{-1}\left(\Sigma_\phi+\lambda I\right)\left(\Sigma_\phi+\lambda I\right)^{-1}-\left(\hat{\Sigma}_\phi+\lambda I\right)^{-1}\left(\hat{\Sigma}_\phi+\lambda I\right)\left(\Sigma_\phi+\lambda I\right)^{-1}\right\|_{\mathcal{L}^\infty(\mathcal{H}_\phi)}\\&=\left\|\hat{\Sigma}_\phi^{-\lambda}\left[\left(\Sigma_\phi+\lambda I\right)-\left(\hat{\Sigma}_\phi+\lambda I\right)\right]\Sigma_\phi^{-\lambda}\right\|_{\mathcal{L}^\infty(\mathcal{H}_\phi)}\\&\leq\left\|\hat{\Sigma}_\phi^{-\lambda}\right\|_{\mathcal{L}^\infty(\mathcal{H}_\phi)}\left\|\Sigma_\phi-\hat{\Sigma}_\phi\right\|_{\mathcal{L}^\infty(\mathcal{H}_\phi)}\left\|\Sigma_\phi^{-\lambda}\right\|_{\mathcal{L}^\infty(\mathcal{H}_\phi)}\\&\leq\frac{1}{\lambda^2}\left\|\Sigma_\phi-\hat{\Sigma}_\phi\right\|_{\mathcal{L}^\infty(\mathcal{H}_\phi)}\leq\frac{1}{\lambda^2}\left\|\Sigma_\phi-\hat{\Sigma}_\phi\right\|_{\mathcal{L}^2(\mathcal{H}_\phi)}.\end{aligned}$$

Similarly, $\left\|\hat{\Sigma}_\psi^{-\lambda}-\Sigma_\psi^{-\lambda}\right\|_{\mathcal{L}^\infty(\mathcal{H}_\psi)}\leq\frac{1}{\lambda^2}\left\|\Sigma_\psi-\hat{\Sigma}_\psi\right\|_{\mathcal{L}^\infty(\mathcal{H}_\psi)}\leq\frac{1}{\lambda^2}\left\|\Sigma_\psi-\hat{\Sigma}_\psi\right\|_{\mathcal{L}^2(\mathcal{H}_\psi)}$.

Therefore, we have that

$$\begin{aligned}
\mathbf{A} =& \left| \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[\left(\hat{A}_{ij,\phi} - \hat{A}_{ij,\psi} \right)^2 - (A_{ij,\phi} - A_{ij,\psi})^2 \right] \right| \\
=& \left| \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[\left(\hat{A}_{ij,\phi} - \hat{A}_{ij,\psi} \right) - (A_{ij,\phi} - A_{ij,\psi}) \right] \left[\left(\hat{A}_{ij,\phi} - \hat{A}_{ij,\psi} \right) + (A_{ij,\phi} - A_{ij,\psi}) \right] \right| \\
\leq& \kappa \left(\left\| \hat{\Sigma}_{\phi}^{-\lambda} - \Sigma_{\phi}^{-\lambda} \right\|_{\mathcal{L}^\infty(\mathcal{H}_\phi)} + \left\| \hat{\Sigma}_{\psi}^{-\lambda} - \Sigma_{\psi}^{-\lambda} \right\|_{\mathcal{L}^\infty(\mathcal{H}_\psi)} \right) \times \left(\frac{2\kappa}{\lambda} + \frac{2\kappa}{\lambda} \right) \\
=& \frac{4\kappa^2}{\lambda} \left[\left\| \hat{\Sigma}_{\phi}^{-\lambda} - \Sigma_{\phi}^{-\lambda} \right\|_{\mathcal{L}^\infty(\mathcal{H}_\phi)} + \left\| \hat{\Sigma}_{\psi}^{-\lambda} - \Sigma_{\psi}^{-\lambda} \right\|_{\mathcal{L}^\infty(\mathcal{H}_\psi)} \right] \\
\leq& \frac{4\kappa^2}{\lambda^3} \left[\left\| \hat{\Sigma}_{\phi} - \Sigma_{\phi} \right\|_{\mathcal{L}^2(\mathcal{H}_\phi)} + \left\| \hat{\Sigma}_{\psi} - \Sigma_{\psi} \right\|_{\mathcal{L}^2(\mathcal{H}_\psi)} \right].
\end{aligned} \tag{32}$$

Let us define $Z_i^\phi = K_\phi(\cdot, X_i) \otimes_{\mathcal{H}_\phi} K_\phi(\cdot, X_i)$. Then, Z_i^ϕ's are i.i.d random variables, $\mathbb{E}(Z_i^\phi) = \Sigma_\phi$ and $\hat{\Sigma}_\phi - \Sigma_\phi = \frac{1}{n} \sum_{i=1}^n \left[Z_i^\phi - \mathbb{E}(Z_i^\phi) \right]$. Similarly, let us define $Z_i^\psi = K_\psi(\cdot, X_i) \otimes_{\mathcal{H}_\psi} K_\psi(\cdot, X_i)$. Then Z_i^ψ's are i.i.d random variables, $\mathbb{E}(Z_i^\psi) = \Sigma_\psi$ and $\hat{\Sigma}_\psi - \Sigma_\psi = \frac{1}{n} \sum_{i=1}^n \left[Z_i^\psi - \mathbb{E}(Z_i^\psi) \right]$.

Note that,

$$\left\| Z_i^\phi \right\|_{\mathcal{L}^2(\mathcal{H}_\phi)} = \sqrt{\left\langle Z_i^\phi, Z_i^\phi \right\rangle_{\mathcal{L}^2(\mathcal{H}_\phi)}} = \langle K_\phi(\cdot, X_i), K_\phi(\cdot, X_i) \rangle_{\mathcal{H}_\phi} = K_\phi(X_i, X_i) \leq \kappa := B.$$

Further,

$$\begin{aligned}
&\mathbb{E} \left\| Z_i^\phi - \mathbb{E}(Z_i^\phi) \right\|_{\mathcal{L}^2(\mathcal{H}_\phi)}^2 = \mathbb{E} \left[\left\langle Z_i^\phi, Z_i^\phi \right\rangle_{\mathcal{L}^2(\mathcal{H}_\phi)} \right] - \langle \Sigma_\phi, \Sigma_\phi \rangle_{\mathcal{L}^2(\mathcal{H}_\phi)} \leq \quad \mathbb{E} \left[\left\langle Z_i^\phi, Z_i^\phi \right\rangle_{\mathcal{L}^2(\mathcal{H}_\phi)} \right] \\
=& \mathbb{E} \left[\langle K_\phi(\cdot, X_i), K_\phi(\cdot, X_i) \rangle_{\mathcal{H}_\phi}^2 \right] = \mathbb{E} \left[K_\phi(X_i, X_i)^2 \right] \leq \kappa^2 := \theta^2.
\end{aligned}$$

Similarly, we can show that $\left\| Z_i^\psi \right\|_{\mathcal{L}^2(\mathcal{H}_\psi)} \leq \kappa = B$ and $\mathbb{E} \left\| Z_i^\phi - \mathbb{E}(Z_i^\phi) \right\|_{\mathcal{L}^2(\mathcal{H}_\phi)}^2 \leq \kappa^2 = \theta^2$.

Note that since $K(\cdot, \cdot)$ is bounded and continuous,$\mathcal{H}_\phi$ and $\mathcal{H}_\psi$ are separable Hilbert spaces. Now, using Bernstein's inequality for separable Hilbert spaces (Theorem D.1 in Sriperumbudur & Sterge (2022)), we have that, for any $0 < \delta < 1$,

$$P \left(\left\| \hat{\Sigma}_\phi - \Sigma_\phi \right\|_{\mathcal{L}^2(\mathcal{H}_\phi)} \geq \frac{2\kappa \log(\frac{6}{\delta})}{n} + \sqrt{\frac{2\kappa^2 \log(\frac{6}{\delta})}{n}} \right) \leq \frac{\delta}{3}$$

and

$$P \left(\left\| \hat{\Sigma}_\psi - \Sigma_\psi \right\|_{\mathcal{L}^2(\mathcal{H}_\phi)} \geq \frac{2\kappa \log(\frac{6}{\delta})}{n} + \sqrt{\frac{2\kappa^2 \log(\frac{6}{\delta})}{n}} \right) \leq \frac{\delta}{3}.$$

Therefore, we have that, for any $0 < \delta < 1$,

$$P \left(\mathbf{A} = \left| \left[\hat{d}_\lambda^{\text{UKP}}(\phi, \psi) \right]^2 - \left[\tilde{d}_\lambda^{\text{UKP}}(\phi, \psi) \right]^2 \right| \geq \frac{8\kappa^2}{\lambda^3} \left[\frac{2\kappa \log(\frac{6}{\delta})}{n} + \sqrt{\frac{2\kappa^2 \log(\frac{6}{\delta})}{n}} \right] \right) \leq \frac{2\delta}{3}.$$

We now proceed to bound $\mathbf{B}$.

Let us define

$$\begin{aligned}
b_{ij} :=& \frac{1}{n^2} \left[\left\langle \Sigma_\phi^{-\frac{\lambda}{2}} K_\phi(\cdot, X_i), \Sigma_\phi^{-\frac{\lambda}{2}} K_\phi(\cdot, X_j) \right\rangle_{\mathcal{H}_\phi} - \left\langle \Sigma_\psi^{-\frac{\lambda}{2}} K_\psi(\cdot, X_i), \Sigma_\psi^{-\frac{\lambda}{2}} K_\psi(\cdot, X_j) \right\rangle_{\mathcal{H}_\psi} \right]^2 \\
=& \frac{1}{n^2} \left[A_{ij,\phi} - A_{ij,\psi} \right]^2.
\end{aligned}$$

Then, clearly, we have that $(b_{ij})_{i,j=1,i\neq j}^{n}$'s are i.i.d random variables. Similarly, $(b_{ii})_{i=1}^{n}$ are i.i.d random variables. Further, $\mathbb{E}(b_{ij}) = \frac{\left[d^{\text{UKP}}_{\lambda,K}(\phi,\psi)\right]^2}{n^2}$ if $i \neq j$ and $|b_{ij}| \leq \frac{1}{n^2}\left[|A_{ij,\phi}| + |A_{ij,\psi}|\right]^2 \leq \frac{4\kappa^2}{\lambda^2 n^2}$ for any i, j. Therefore, $|\mathbb{E}(b_{ij})| \leq \mathbb{E}\,|b_{ij}| \leq \frac{4\kappa^2}{\lambda^2 n^2}$ for any i, j and $\left[d^{\text{UKP}}_{\lambda,K}(\phi,\psi)\right]^2 \leq \frac{4\kappa^2}{\lambda^2}$.

Now, we have that,

$$\mathbb{E}\left[\tilde{d}^{\text{UKP}}_{\lambda}(\phi,\psi)\right]^2 = \frac{n(n-1)}{n^2}\left[d^{\text{UKP}}_{\lambda,K}(\phi,\psi)\right]^2 + n\mathbb{E}(b_{11}).$$

Consequently, $\left[d^{\text{UKP}}_{\lambda,K}(\phi,\psi)\right]^2 - \mathbb{E}\left[\tilde{d}^{\text{UKP}}_{\lambda}(\phi,\psi)\right]^2 = \frac{1}{n}\left[d^{\text{UKP}}_{\lambda,K}(\phi,\psi)\right]^2 - n\mathbb{E}(b_{11})$. Therefore,

$$\left|\left[d^{\text{UKP}}_{\lambda,K}(\phi,\psi)\right]^2 - \mathbb{E}\left[\tilde{d}^{\text{UKP}}_{\lambda}(\phi,\psi)\right]^2\right| \leq \frac{8\kappa^2}{\lambda^2 n}.$$

Now, using McDiarmid's inequality, we have that,

$$P\left(\left|\left[\tilde{d}^{\text{UKP}}_{\lambda}(\phi,\psi)\right]^2 - \mathbb{E}\left[\tilde{d}^{\text{UKP}}_{\lambda}(\phi,\psi)\right]^2\right| \geq \frac{4\kappa^2}{\lambda^2}\sqrt{\frac{2\log(\frac{6}{\delta})}{n}}\right) \leq \frac{\delta}{3}.$$

Therefore, we have that,

$$P\left(\mathbf{B} \geq \frac{\kappa^2}{\lambda^2}\left[\frac{8}{n} + 4\sqrt{\frac{2\log(\frac{6}{\delta})}{n}}\right]\right) \leq \frac{\delta}{3}.$$

Finally, we have that,

$$P\left(\mathbf{A}+\mathbf{B} \leq \frac{8\kappa^3}{\lambda^3}\left[\frac{2\log(\frac{6}{\delta})}{n} + \sqrt{\frac{2\log(\frac{6}{\delta})}{n}}\right] + \frac{4\kappa^2}{\lambda^2}\left[\frac{2}{n} + \sqrt{\frac{2\log(\frac{6}{\delta})}{n}}\right]\right) \geq 1-\delta,$$

which completes the proof. □

A.8 Bound on the sensitivity of risk functional for a pair of representations

Theorem 12. *Assume $\lambda > 0$ and consider a spectral regularizer g_λ that satisfies Assumptions (A_1), (A_2), (A_3) and (A_4). Let $\mathcal{R}^{g_\lambda}_{\varphi}(\eta) = \mathbb{E}\left\{[\alpha^{\varphi}_{g_\lambda}(X) - \eta(X)]^2\right\}$ be the squared-loss-regression based risk functional corresponding to the kernel regularized population estimator of the regression function $\eta = \mathbb{E}(Y|X)$ using the pullback kernel K_φ and the spectral regularizer g_λ corresponding to any representation $\varphi : \mathbb{R}^d \to \mathbb{R}^{out}$. Then, for any two given representations $\phi : \mathbb{R}^d \to \mathbb{R}^k$ and $\psi : \mathbb{R}^d \to \mathbb{R}^l$ with corresponding base kernel K, the sensitivity of the risk functional can be bounded in terms of the UKP distance $d^{\text{UKP}}_{g_\lambda,K,\mathcal{L}^\infty}$ as follows -*

$$|\mathcal{R}^{g_\lambda}_{\phi}(\eta) - \mathcal{R}^{g_\lambda}_{\psi}(\eta)| \leq 4 \times d^{\text{UKP}}_{g_\lambda,K,\mathcal{L}^\infty}(\phi,\psi) \times \|\eta\|^2_{L^2(P_X)} \tag{33}$$

In particular, when we choose g_λ to be the Tikhonov regularizer, the risk sensitivity bound can be expressed as-

$$|\mathcal{R}^{\lambda}_{\phi}(\eta) - \mathcal{R}^{\lambda}_{\psi}(\eta)| \leq 4 \times d^{UKP}_{\lambda,K,\mathcal{L}^\infty}(\phi,\psi)(\phi,\psi) \times \|\eta\|^2_{L^2(P_X)} \tag{34}$$

Proof. Observe that,

$$\begin{aligned}\mathcal{R}_\phi(\eta) =& \mathbb{E}\left\{[\alpha^\phi_{g_\lambda}(X) - \eta(X)]^2\right\}\\ =& \left\|(\mathfrak{I}_\phi g_\lambda(\Sigma_\phi)\mathfrak{I}^*_\phi - I)\eta\right\|^2_{L^2(P_X)}\\ =& \left\|(\mathfrak{I}_\phi \mathfrak{I}^*_\phi g_\lambda(\mathcal{T}_\phi) - I)\eta\right\|^2_{L^2(P_X)}\\ =& \|(\mathcal{T}_\phi g_\lambda(\mathcal{T}_\phi) - \mathfrak{I})\eta\|^2_{L^2(P_X)}\\ =& \left\langle (\mathcal{T}_\phi g_\lambda(\mathcal{T}_\phi) - I)^2 \eta, \eta\right\rangle_{L^2(P_X)}\\ =& \| (\mathcal{T}_\phi g_\lambda(\mathcal{T}_\phi) - I)\, \eta\|^2_{L^2(P_X)}\\ =& \|u\|^2\end{aligned}$$

where $u := (\mathcal{T}_\phi g_\lambda(\mathcal{T}_\phi) - I)\,\eta$.

Similarly,

$$\begin{aligned}\mathcal{R}_\psi(\eta) =& \mathbb{E}\left\{[\alpha^\psi_{g_\lambda}(X) - \eta(X)]^2\right\}\\ =& \left\langle (\mathcal{T}_\psi g_\lambda(\mathcal{T}_\psi) - I)^2 \eta, \eta\right\rangle_{L^2(P_X)}\\ =& \| (\mathcal{T}_\psi g_\lambda(\mathcal{T}_\psi) - I)\, \eta\|^2_{L^2(P_X)}\\ =& \|v\|^2\end{aligned}$$

where $v := (\mathcal{T}_\psi g_\lambda(\mathcal{T}_\psi) - I)\,\eta$.

Therefore, we have that

$$\begin{aligned}&|\mathcal{R}_\phi(\eta) - \mathcal{R}_\psi(\eta)|\\ &= |\|u\|^2 - \|v\|^2|\\ &= |\langle u, u\rangle - \langle v, v\rangle|\\ &= |\langle u, u\rangle + \langle u, v\rangle - \langle u, v\rangle - \langle v, v\rangle|\\ &= |\langle u + v, u\rangle - \langle u + v, v\rangle|\\ &= |\langle u + v, u - v\rangle|\\ &= |\langle (\mathcal{T}_\phi g_\lambda(\mathcal{T}_\phi) + \mathcal{T}_\psi g_\lambda(\mathcal{T}_\psi) - 2I)\,\eta, (\mathcal{T}_\phi g_\lambda(\mathcal{T}_\phi) - \mathcal{T}_\psi g_\lambda(\mathcal{T}_\psi))\,\eta\rangle_{L^2(P_X)}|\\ &\overset{(i)}{\leq} \| (\mathcal{T}_\phi g_\lambda(\mathcal{T}_\phi) + \mathcal{T}_\psi g_\lambda(\mathcal{T}_\psi) - 2I)\,\eta\|_{L^2(P_X)} \| (\mathcal{T}_\phi g_\lambda(\mathcal{T}_\phi) - \mathcal{T}_\psi g_\lambda(\mathcal{T}_\psi))\,\eta\|_{L^2(P_X)}\\ &\overset{(ii)}{\leq} \|(\mathcal{T}_\phi g_\lambda(\mathcal{T}_\phi) + \mathcal{T}_\psi g_\lambda(\mathcal{T}_\psi) - 2I)\|_{L^\infty(L^2(P_X))}\|\eta\|_{L^2(P_X)}\|\\ &\times (\mathcal{T}_\phi g_\lambda(\mathcal{T}_\phi) - \mathcal{T}_\psi g_\lambda(\mathcal{T}_\psi))\|_{L^\infty(L^2(P_X))}\|\eta\|_{L^2(P_X)}\\ &\leq \left\{\|\mathcal{T}_\phi g_\lambda(\mathcal{T}_\phi)\|_{L^\infty(L^2(P_X))} + \|\mathcal{T}_\psi g_\lambda(\mathcal{T}_\psi)\|_{L^\infty(L^2(P_X))} + 2\right\} \times d^{UKP}_{g_\lambda,K,\mathcal{L}^\infty}(\phi,\psi) \times \|\eta\|^2_{L^2(P_X)}\\ &\leq (1+1+2) \times d^{UKP}_{g_\lambda,K,\mathcal{L}^\infty}(\phi,\psi) \times \|\eta\|^2_{L^2(P_X)}\\ &= 4 \times d^{UKP}_{g_\lambda,K,\mathcal{L}^\infty}(\phi,\psi) \times \|\eta\|^2_{L^2(P_X)},\end{aligned}$$

where (i) follows from Cauchy-Schwarz inequality and (ii) follows from the definition of operator norms. □

A.9 The integrated UKP pseudometric

Let (Λ, μ) be a σ-finite measure space (typically $\Lambda \subset (0, \infty)$). Let $w : \Lambda \to [0, \infty)$ be a measurable weight function. Define the integrated UKP (squared) pseudometric-

$$\mathcal{D}^2_{w,K}(\phi,\psi) := \int_\Lambda w(\lambda)\left[d^{UKP}_{g_\lambda,K,\mathcal{L}^p}(\phi,\psi)\right]^2 \mu(d\lambda)$$

Its corresponding estimator is the λ-integral of the closed-form function of the data-

$$\widehat{\mathcal{D}}^2_{w,K}(\phi,\psi) := \int_\Lambda w(\lambda)\left[\widehat{d^{UKP}_{g_\lambda,K,\mathcal{L}^p}}(\phi,\psi)\right]^2 \mu(d\lambda).$$

Assuming, $J_2 := \int_\Lambda \frac{w(\lambda)}{\lambda^2}\mu(d\lambda) < \infty, \quad J_3 := \int_\Lambda \frac{w(\lambda)}{\lambda^3}\mu(d\lambda) < \infty$, $\mathcal{D}^2_{w,K}(\phi,\psi)$ forms a valid pseudometric and we can obtain similar results as in UKP .

A.10 Computational complexity of $\hat{d}^{UKP}_{\lambda,K,\mathcal{L}^2}$

From the expression of the estimator $\hat{d}^{UKP}_{\lambda,K,\mathcal{L}^2}$ in Proposition 5, it can be shown that its computational complexity is $O(n^3)$, where n is the sample size. Notably, the GULP distance proposed in Boix-Adsera et al. (2022) shares the same complexity. The primary computational cost arises from inverting the Gram matrix, which can be reduced using kernel approximation techniques like Random Fourier Features (RFF) or Nyström approximation. For example, by using D RFF samples from the spectral distribution of the kernel K or D subsamples from the n data samples in the Nyström method, the complexity of the UKP distance estimator $\hat{d}^{UKP}_{\lambda,K,\mathcal{L}^2}(\phi,\psi)$ can be reduced from $O(n^3)$ to $O(nD^2 + D^3)$, which is significantly lower than $O(n^3)$ when $D \ll n$. Exploring the tradeoff between the statistical accuracy of UKP distance estimation and the computational efficiency of kernel approximation methods is a promising direction for future research.

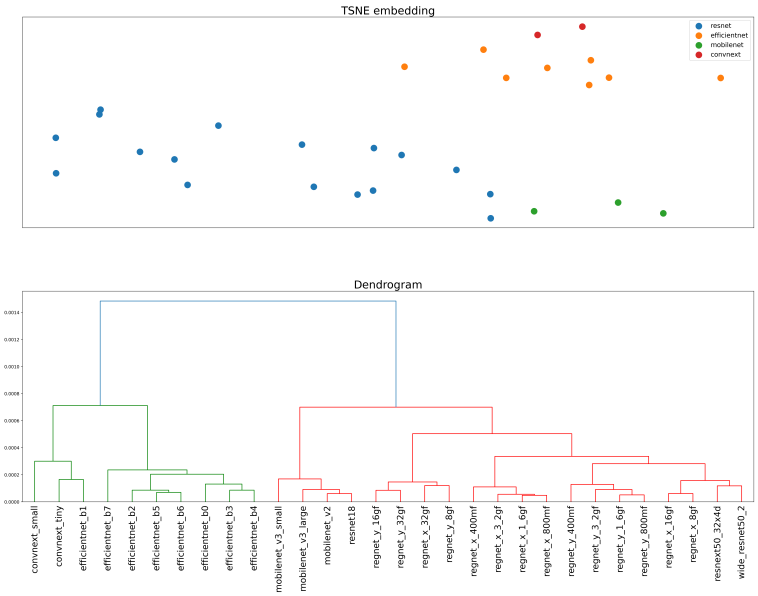


Figure 2: Clustering based on UKP distance is sensitive to differences in architectures of neural network models.

B Appendix: Additional Experiments

In this appendix, we discuss the ability of UKP to identify differences in architectures and inductive biases and provide additional experimental results.

B.1 Ability of UKP to identify differences in architectures and inductive biases

A key source of inductive biases in neural network models is their architecture, with features such as residual connections and variations in convolutional filter complexity shaping the representations learned during training. As a pseudometric over feature space, the UKP distance is expected to capture intrinsic differences in these inductive biases, which are known to impact generalization performance across tasks. To explore this, we analyze representations from 35 pre-trained neural network architectures used for image classification, described in detail in Section B.3 of the Appendix.

We estimate pairwise UKP distances between model representations using 3,000 images from the validation set of the ImageNet dataset (Krizhevsky et al., 2012), a regularization parameter $\lambda = 1$ and a Gaussian kernel with bandwidth $\sigma = 10$. The tSNE embedding method is then used to embed these representations into 2-D space utilizing the distance measures given by the UKP pseudometric. Concurrently, we perform an agglomerative (bottom-up) hierarchical clustering of the representations based on the pairwise UKP distances and obtain the corresponding dendrogram. We observe in Fig. 2 that similar architectures which share important properties, such as the Regnets and Resnets are clustered together, while they are well separated from smaller efficient architectures such as MobileNets and ConvNexts. This demonstrates that the UKP distance effectively captures notions of similarity and dissimilarity aligned with interpretable notions based on inductive biases. Further comparisons with baseline measures, such as GULP and CKA, presented in Fig. 9 in Section B.3 of the Appendix demonstrate that UKP often provides superior clustering quality. We would like to note here that the choice of the kernel function for the UKP pseudometric should be driven by the nature of inductive bias that will be useful for the tasks for which the representations/features of interest will be used. Additional discussion regarding kernel (and kernel parameter) selection is provide in Section B.3 of the Appendix.

B.2 MNIST EXPERIMENTS

Training details We have already described the architectures of the 50 ReLU networks we trained for experiments using the MNIST dataset in Section 5.1. We used the uniform Kaiming initialization He et al. (2015) for initializing the network weights for every network with a specific width and depth, while the biases are set to zero at initialization. We used a single A100 GPU on the Google Colab platform. We chose to use the Adam optimizer with a learning rate of 10^{-4} and a batch size of 100 to train the 50 ReLU networks. We follow a training scheme similar to that used in Boix-Adsera et al. (2022).

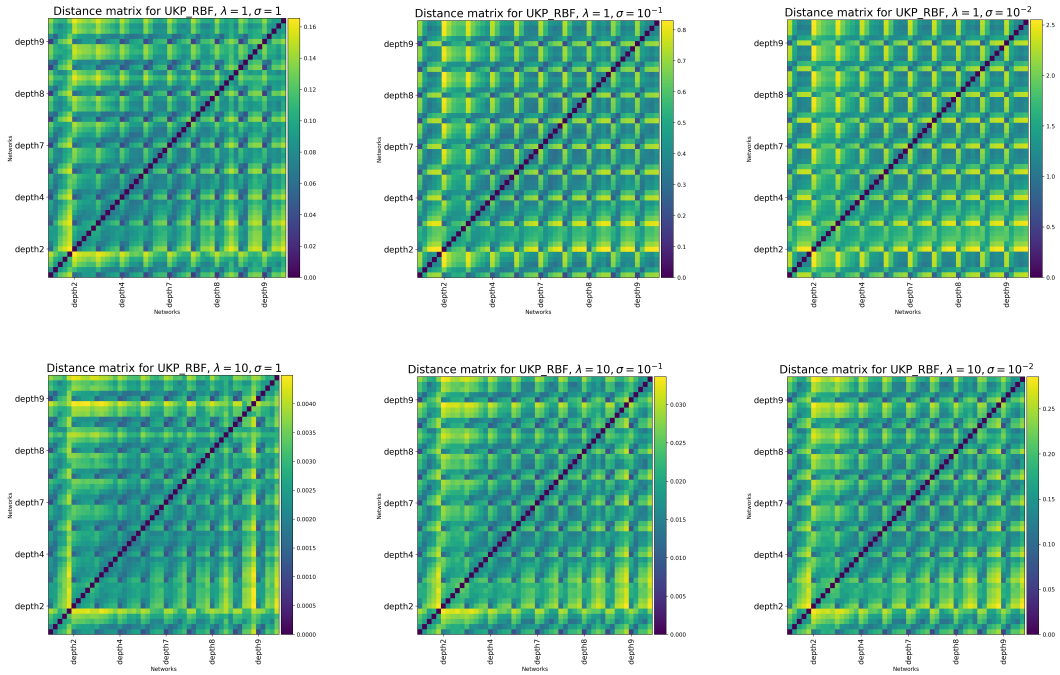


Figure 3: Heatmaps representing UKP distance between pairs of fully-connected ReLU networks of different depths and widths. We choose the kernel for the UKP distance to be the Gaussian RBF kernel with bandwidth $\sigma \in \{1, 10^{-1}, 10^{-2}\}$ along with the regularization parameter $\lambda \in \{1, 10\}$. Along the rows and columns of each of the heatmaps, the ReLU networks are first arranged in order of increasing depth, and then in order of increasing width inside each specific depth level. Darker colors indicate smaller value of UKP distance according to the scale attached to each heatmap.

Clustering of representations based on UKP aligns with architectural characteristics of networks We observe in Fig. 3 that a repeating block structure emerges in each heatmap, with each block corresponding to networks with the same depth. Within each block, i.e., same depth, the pairwise similarities between networks with different widths are higher if the difference of widths of the pair of networks is small, and the similarities are lower otherwise. Further, it seems that the relative difference between networks with different depths is amplified (in terms of the UKP distance) if the depths of the networks are larger. For e.g. the contrast between a width 500 and width 600 network is higher when the depth is 9 for both networks, compared to the scenario where both networks have depth 2. We also perform an agglomerative (bottom-up) hierarchical clustering of the representations based on the pairwise UKP distances and obtain the corresponding dendrograms as shown in Fig. 4. The dendrograms also exhibit separation between deeper networks (depths 7,8 and 9) and shallow networks (depths 2,4 and 6) over a range of (λ, σ) choices for the UKP distance with Gaussian RBF kernel. This indicates that the UKP distance is able to capture the relevant differences in predictive performance that are induced by architectural differences in these networks, over a wide range of values of its tuning parameters.

Generalization ability on kernel ridge regression tasks We consider the same setup as discussed in Section 5.1. Supplementing our choices of $\lambda = 10^{-2}$ and $\sigma = 10^{-1}$ corresponding to synthetic

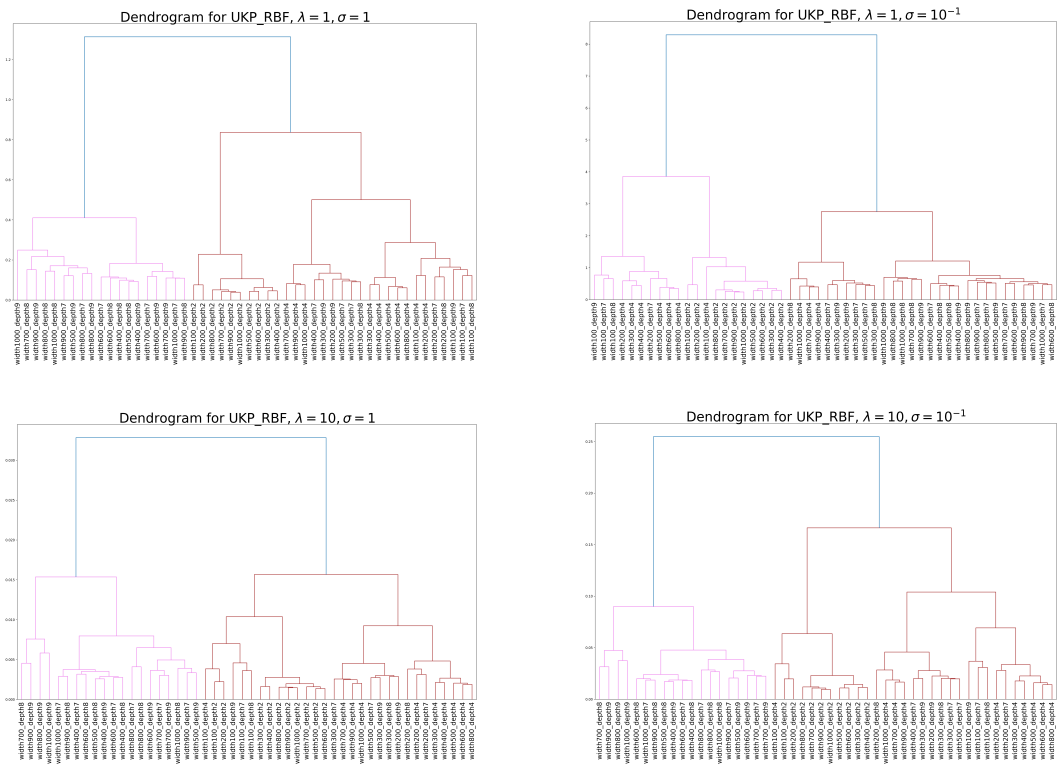


Figure 4: Dendrograms corresponding to agglomerative hierarchical clustering of representations of 50 ReLU networks based on UKP distance

kernel ridge regression tasks with Gaussian RBF kernel, we now consider $\lambda \in \{10^{-2}, 1\}$ and $\sigma \in \{10^{-1}, 1\}$. In Fig. 5, we plot the Spearman's ρ rank correlation coefficient between the $err_{\phi,\psi}$'s as defined in Section 5.1 and the pairwise distances between the representations using the following distances - CCA, linear CKA, nonlinear CKA with Gaussian RBF kernel, GULP and UKP with Gaussian RBF kernel.

When $(\lambda = 10^{-2}, \sigma = 10^{-1})$ and $(\lambda = 1, \sigma = 10^{-1})$, we observe from Fig. 5 that the pairwise UKP distance is positively correlated to a moderate extent with the collection of $err_{\phi,\psi}$'s, as evident from the large positive values of the blue bars. In contrast, GULP distances show inconsistent behavior across different levels of regularization, while CCA and linear CKA distances show a much lower positive correlation with generalization performance (with CCA even showing negative correlation when $(\lambda = 1, \sigma = 10^{-1})$). For the remaining choices, none of the distance measures show any consistent behavior, which indicates that an increase in the number of samples used to approximate the model representations may improve the performance of these distance measures.

Unsurprisingly, as a consequence of the relationship between CKA and UKP , as discussed in Section 4.1, the performance of the CKA distance, when using the Gaussian RBF kernel (with the corresponding bars shown in red), is comparable to that of UKP with the same choice of kernel. This similarity in the information conveyed by these two measures can be empirically observed through their scatterplots and the Pearson product-moment correlation coefficient under various choices of tuning parameters. As shown in Fig. 6, the nearly linear positive relationship between UKP and CKA distances, when both are used with a Gaussian RBF kernel, along with the high positive correlation coefficient, suggests that either measure could be effectively used in practice for comparing representations. However, the UKP distance may be preferred over the CKA distance due to its pseudometric properties, particularly the triangle inequality, which proves to be especially useful. In contrast, CKA, being a measure akin to a normalized inner product bounded between 0 and 1, does not satisfy the properties of a pseudometric and may lead to misleading intuitions when comparing different representations.

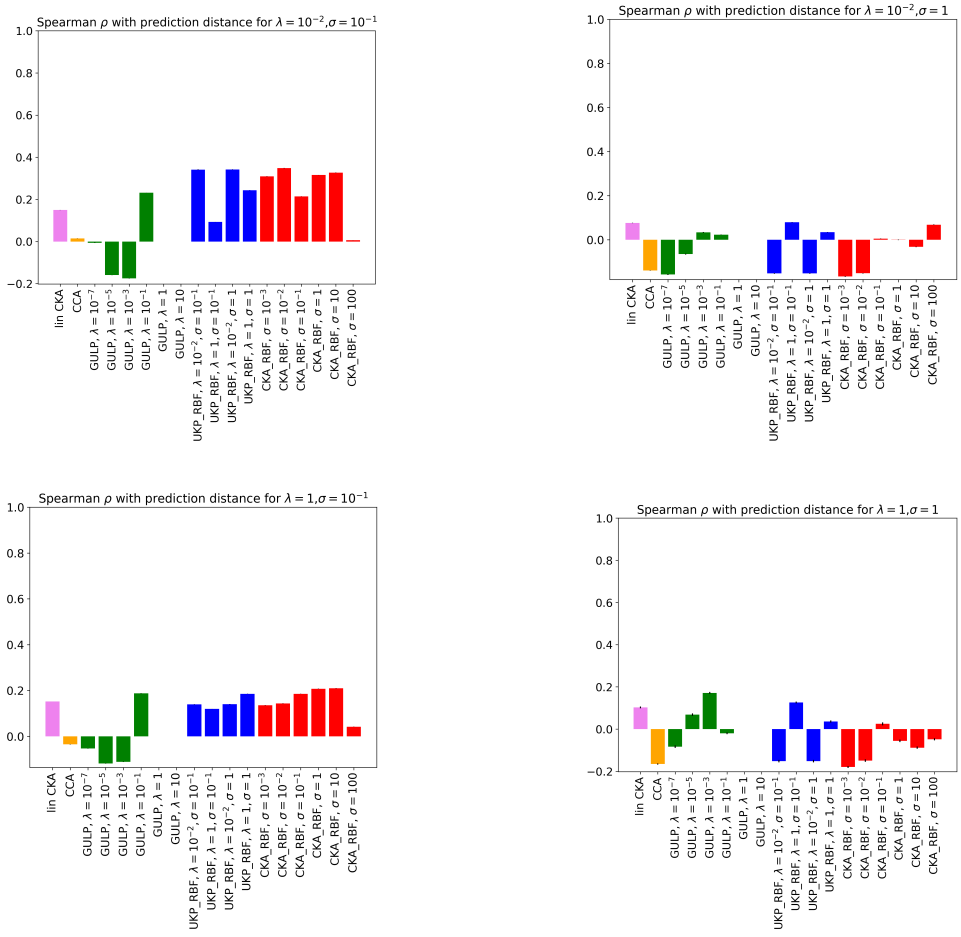


Figure 5: Spearman's ρ rank correlation coefficient between generalization of kernel ridge regression-based predictors with various distance measures between representations. We report the average correlation across 10 random synthetic kernel ridge regression tasks. Results are similar for 30 trials. Error bars are negligibly small and hence not visible.

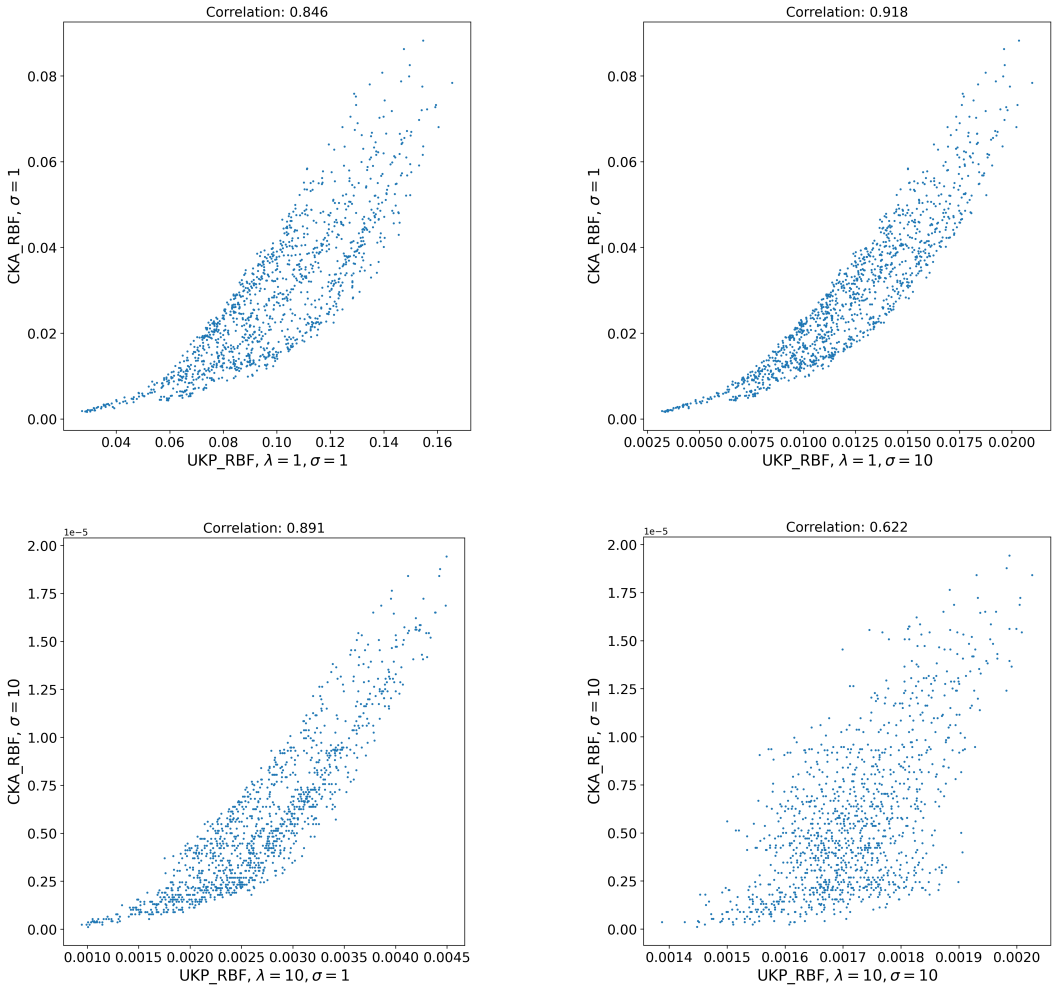


Figure 6: Correlation plots between UKP and CKA measures with Gaussian RBF kernel between $\binom{50}{2}$ pairs of ReLU networks trained on MNIST data. Plot titles display the Pearson product-moment correlation coefficient between the distance measures on the two axes.

B.3 IMAGENET EXPERIMENTS

Architectures used and data description In our experiments, we utilized 35 pretrained models known for achieving state-of-the-art (SOTA) performance in the ImageNet Object Localization Challenge on Kaggle Howard et al. (2018), available from PyTorch (2024). These models are categorized based on their architectural types as follows:

- **ResNets (17 models):** regnet_x_16gf, regnet_x_1_6gf, regnet_x_32gf, regnet_x_3_2gf, regnet_x_400mf, regnet_x_800mf, regnet_x_8gf, regnet_y_16gf, regnet_y_1_6gf, regnet_y_32gf, regnet_y_3_2gf, regnet_y_400mf, regnet_y_800mf, regnet_y_8gf, resnet18, resnext50_32x4d, wide_resnet50_2
- **EfficientNets (8 models):** efficientnet_b0, efficientnet_b1, efficientnet_b2, efficientnet_b3, efficientnet_b4, efficientnet_b5, efficientnet_b6, efficientnet_b7
- **MobileNets (3 models):** mobilenet_v2, mobilenet_v3_large, mobilenet_v3_small
- **ConvNeXts (2 models):** convnext_small, convnext_tiny
- **Other Architectures (5 models):** alexnet, googlenet, inception, mnasnet, vgg16 .

The penultimate layer dimensions for these networks, corresponding to the representation sizes, vary from 400 to 4096 depending on the architecture. Each model processes input data as 3-channel RGB images, with each channel having dimensions of 224 × 224 pixels. To approximate the model representations learned by these models using finite-dimensional representations, we used 3000 images from the validation set of the ImageNet dataset. These images were normalized with a mean of (0.485, 0.456, 0.406) and a standard deviation of (0.229, 0.224, 0.225) for each RGB channel. Our choice of models and input preprocessing parameters is similar to those used in Boix-Adsera et al. (2022).

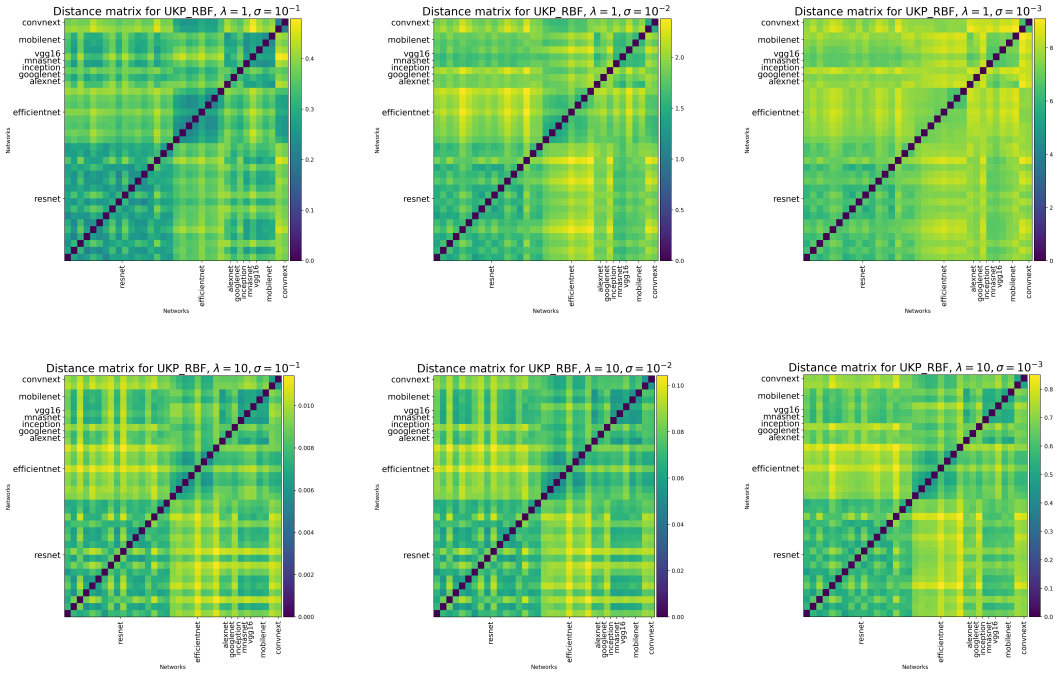


Figure 7: Heatmaps representing UKP distance between pairs of networks of different architecture, pretrained on ImageNet data. We choose the kernel for the UKP distance to be the Gaussian RBF kernel with bandwidth $\sigma \in \{10^{-1}, 10^{-2}, 10^{-3}\}$ along with the regularization parameter $\lambda \in \{1, 10\}$. Along the rows and columns of each of the heatmaps, the networks are arranged in the following order from left to right and top to bottom - ResNets, EfficientNets, Other Architectures, MobileNets and ConvNexts. Darker colors indicate smaller value of UKP distance according to the scale attached to each heatmap.

Clustering of representations based on UKP aligns with architectural characteristics of networks We are interested in observing whether the UKP pseudometric is capable of capturing intrinsic differences in predictive performances of different representations. Such intrinsic differences are often the result of the different inductive biases we encode into networks through the choice of architectures, among other factors.

We first discuss the main architectural similarities and differences between ResNet, RegNet, EfficientNet, MobileNet, alexnet, googlenet, inception, mnasnet, and vgg16, which are controlled by how they address depth, efficiency, and feature extraction. Alexnet and vgg16 are older architectures that use standard convolutional layers arranged in sequential blocks, with vgg16 deepening the network significantly compared to alexnet. Googlenet introduced Inception modules, which combine multiple convolution filters of different sizes to capture multi-scale features, making it more efficient than alexNet and vgg16. Different Inception architectures have been built using the Inception module of Googlenet. ResNet brought the innovation of residual connections (skip connections) to address the vanishing gradient problem, enabling very deep networks, while RegNet refined this concept by creating more regular, scalable structures without explicit skip connections. EfficientNet and mnasnet focus on balanced scaling (depth, width, resolution) and use of MBConv blocks for efficiency, with EfficientNet employing a compound scaling formula. MobileNet, like mnasnet, emphasizes depthwise separable convolutions for lightweight, efficient models suitable for mobile devices. In terms of architectural similarities, resNet and regNet share a focus on structured deep architectures, while EfficientNet and MobileNet share efficiency-driven designs for varied hardware constraints. Alexnet, vgg16, and googlenet represent early convolutional architectures, with googlenet's Inception modules providing a bridge to more modern designs. In contrast, vgg16 and ResNet are quite different, with vgg16 being sequential and deep, and ResNet leveraging residual connections.

We observe in Fig. 7 that a block structure emerges in the heatmaps across different choices of the tuning parameters for the UKP distance, especially corresponding to the 4 major groups of architectures ResNets, EfficientNets, MobileNets and ConvNeXts. We also perform an agglomerative (bottom-up) hierarchical clustering of the representations based on the pairwise UKP distances and obtain the corresponding dendrograms as shown in Fig. 8. The dendrograms exhibit a clear separation between the ResNets/RegNets and the remaining architectures over a range of (λ, σ) choices for the UKP distance with Gaussian RBF kernel. This indicates that, for the class of pretrained ImageNet models we consider, the UKP distance captures the relevant differences in predictive performance that are induced by architectural differences in these networks, over a wide range of values of its tuning parameters.

To illustrate that the performance of the UKP pseudometric is reasonably robust to the choice of the regularization parameter λ and kernel parameters (such as bandwidth parameter σ for the Gaussian RBF kernel), we have compared the performance of UKP 's performance with other popular baseline measures such as GULP and CKA. As observed from Fig. 9, the separation between the different classes of networks is more pronounced in the case of UKP than GULP. Additionally, the clustering behaviour within the primary classes of networks is much weaker for the CKA compared to the UKP and GULP measures, and the separation between the different classes is not clear in the case of CKA.

Relationship between UKP and CKA measures The MNIST experiments, along with the theoretical analysis in section 4.1, reveal a similarity between the information conveyed by the UKP and CKA measures when both use the same kernel. This similarity is also empirically confirmed in the ImageNet experiments, as demonstrated by their scatterplots and the Pearson correlation coefficient across different tuning parameters. As illustrated in Fig. 10, there is an almost linear positive relationship between UKP and CKA distances when both utilize a Gaussian RBF kernel. The strong positive correlation suggests that either measure could be effectively used for comparing representations. However, as previously discussed in Section 4.1, UKP may be preferred over CKA due to its pseudometric properties, particularly the triangle inequality, which is especially advantageous. In contrast, CKA, being a measure similar to a normalized inner product bounded between 0 and 1, does not satisfy pseudometric properties and may lead to misleading interpretations when comparing different representations.

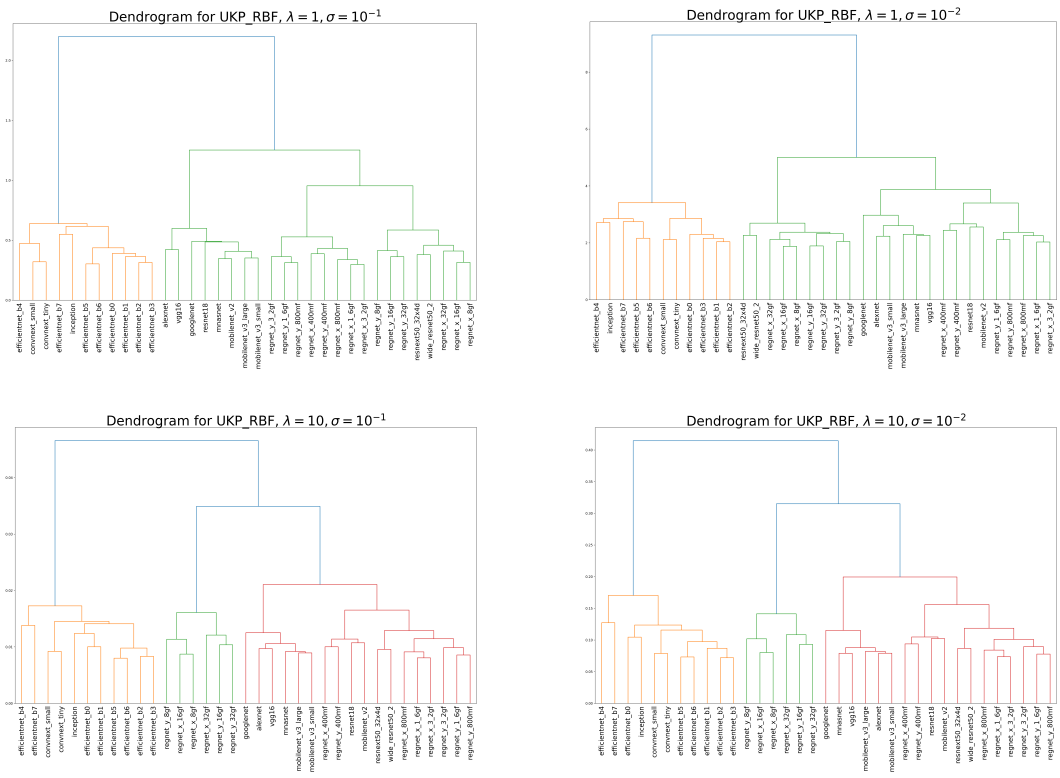
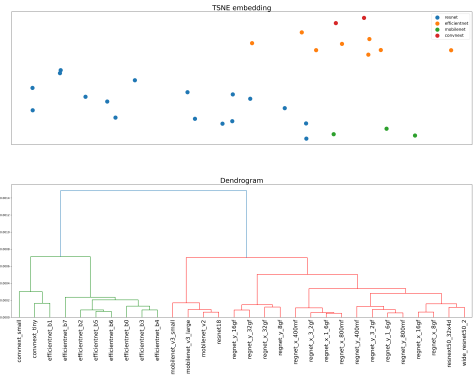


Figure 8: Dendrograms corresponding to agglomerative hierarchical clustering of representations of 35 pretrained ImageNet networks based on UKP distance

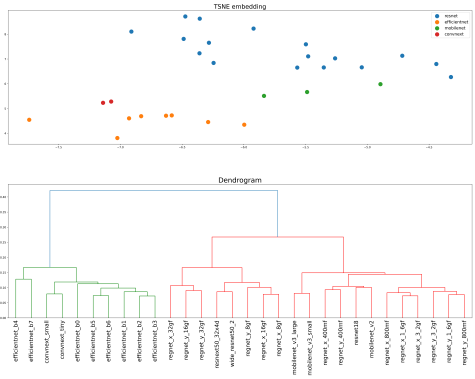
Choice of kernel function The choice of kernel function for the UKP pseudometric should be guided by the inductive bias most relevant to the tasks for which the representations or features of interest will be used. For instance, consider an image classification task where the model's predictions should remain unaffected by image rotations and translations. In this case, we can incorporate this inductive bias into the UKP pseudometric by selecting a rotationally and translationally invariant kernel, such as the Gaussian RBF kernel, as the kernel function for UKP. This approach is particularly useful for comparing the generalization performance of two representations: one obtained through a training or optimization procedure that explicitly enforces rotational and translational invariance, and another trained without such constraints.

Furthermore, even when the true inductive bias is unknown, probing the nature of representations encoded by different models can still provide valuable insights. In this context, the terms "well-specified" and "misspecified" kernels refer, respectively, to choices of kernels for the UKP pseudometric that either capture or fail to capture the required inductive bias for a specific class of downstream tasks utilizing the representations or features of interest. Each kernel choice can be viewed as a selection of particular characteristics of the representations that we aim to investigate.

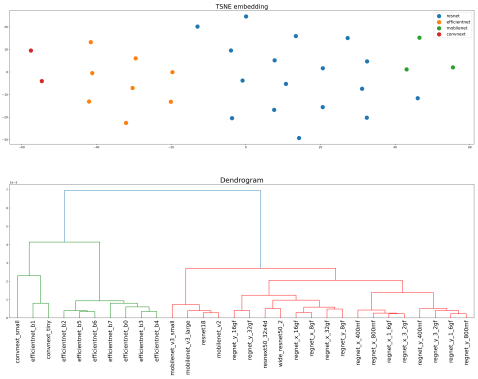
If we have a set of characteristics in mind that we wish to probe, we should select a corresponding set of kernels whose feature maps encode some or all of those characteristics and then analyze the conclusions drawn from using each kernel as the kernel function for the UKP pseudometric. When the kernels are "well-specified", clustering representations based on UKP values can help identify useful pairs of representations for specific downstream tasks. In contrast, when the kernels are "misspecified", the UKP values may still cluster representations with characteristics aligned with the feature maps of the "misspecified" kernels. However, in such cases, the clustering will not be informative for studying generalization performance on downstream tasks. Nonetheless, even with "misspecified kernels", the UKP pseudometric can still provide insights into the characteristics of the representations, though its values will not reliably indicate generalization performance.



(a) UKP (with RBF kernel, regularization parameter $\lambda = 1$ and kernel bandwidth $\sigma = 10$)



(b) GULP (with regularization parameter $\lambda = 1$)



(c) CKA (with RBF kernel and kernel bandwidth $\sigma = 10$)

Figure 9: tSNE embeddings and dendrograms corresponding to agglomerative hierarchical clustering of representations of 35 pretrained ImageNet networks based on UKP (with Gaussian RBF kernel, regularization parameter $\lambda = 1$ and kernel bandwidth $\sigma = 10$), GULP (with regularization parameter $\lambda = 1$) and CKA (with Gaussian RBF kernel and kernel bandwidth $\sigma = 10$) distance

Cross-validation or selecting an "optimal" value for the kernel parameters is not necessary in the context of this paper, as our focus is on an exploratory comparison of the inductive biases encoded by different representations. For example, consider a scenario where we hypothesize that rotational and/or translational invariance are the key inductive biases required for good generalization performance, as in image classification tasks. In this case, the Gaussian RBF kernel is a natural choice. Since the Gaussian RBF kernel remains rotationally and translationally invariant for any value of its bandwidth parameter—which controls the "scale" at which the kernel perceives the representations—the UKP pseudometric should, in principle, capture the extent to which different representations encode rotational and translational invariance, regardless of the specific choice of bandwidth.

Of course, no experimental setup is ever exhaustive. In our study, we focus on datasets from the image domain (MNIST and ImageNet) to illustrate two of the simplest and most fundamental invariances -rotational and translational invariance - which are relevant to most image-related tasks. This consideration motivated our choice of the Gaussian RBF kernel as the kernel function for the UKP pseudometric in our experiments.

Code implementation The Python code for running all the experiments in this paper is available in the following Anonymous GitHub repository: `https://anonymous.4open.science/r/Uniform-Kernel-Prober-ICLR-2026-20792/`. The code for comparing our proposed

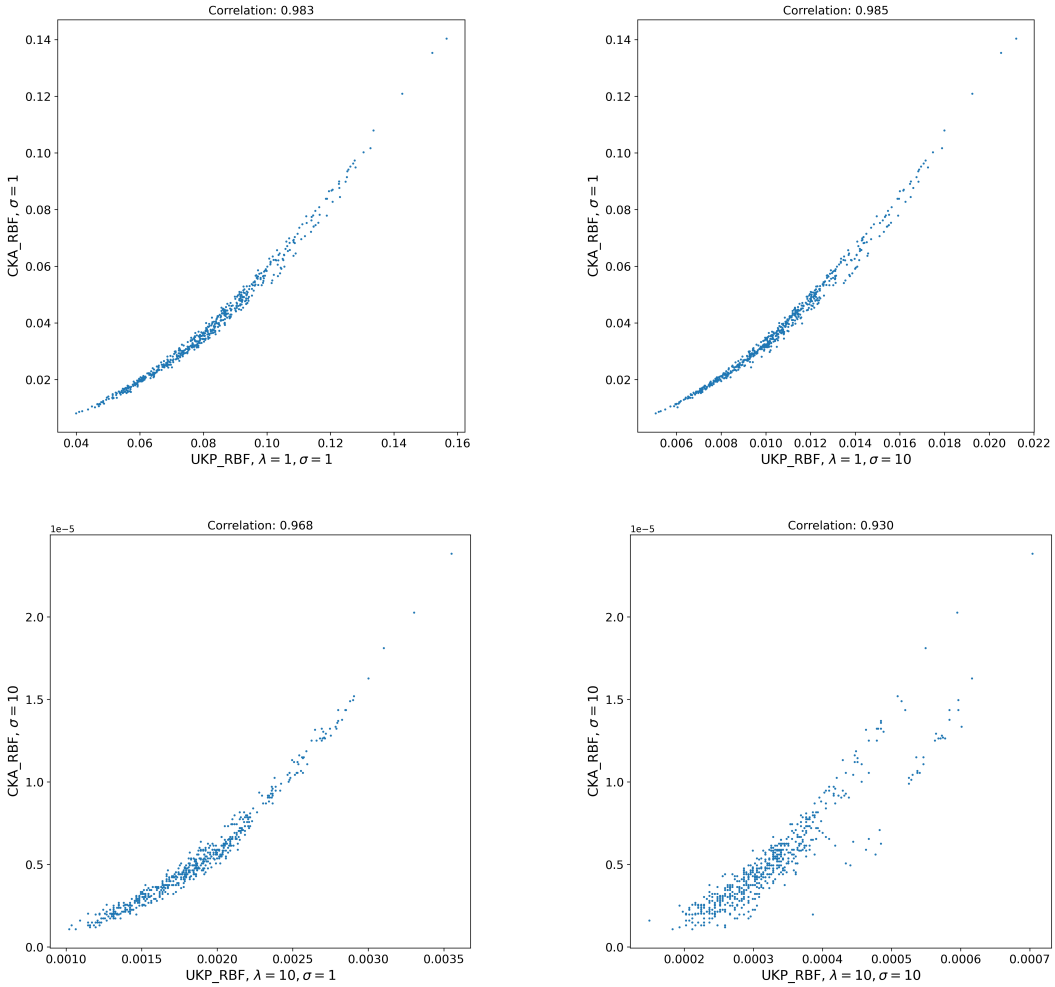


Figure 10: Correlation plots between UKP and CKA measures with Gaussian RBF kernel between $\binom{35}{2}$ pairs of networks with different architectures trained on ImageNet data. Plot titles display the Pearson product-moment correlation coefficient between the distance measures on the two axes.

UKP pseudometric to other distance measures has been adapted from `https://github.com/sgstepaniants/GULP`.