Maximum Mean Discrepancy Tests via Learned Data-Dependent Encoder Kernels

Winston Yu
Advisor: Alex Cloninger

June 2023

Abstract

Motivated by the dimensionality reduction abilities of encoder neural networks, we present a new hypothesis test for high-dimensional data, such as images, for which a low-dimensional latent space is an appropriate choice. An approximation with similar capability to the proposed test using an anisotropic kernel is available. We also describe the test's behavior as the number of datapoints approaches infinity and relate its spectrum to that of the anisotropic kernel by proving a more general fact about the spectra of two kernel integral operators when those kernels are uniformly close. By experiments on image data, we validate our constructions' abilities to distinguish images of different labels.

1 Introduction

The conventional paradigm of computer programming requires that the programmer prescribe actions, expressed to the computer as a sequence of simple tasks, for every possible case that the computer may encounter. For example, matrix multiplication in its simplest form merely asks the computer to calculate the dot product between the rows of one matrix and the columns of another. This paradigm has enabled the automation of tasks whose main effort lies in the sheer number of simple sub-tasks that must be performed. However, automation of other tasks, such as image classification and text generation, is much more difficult, because although humans find their abilities in these fields to be perfectly natural, explicitly describing how humans perform these tasks, as the conventional paradigm of programming requires, is almost impossible. Machine learning provides a solution; rather than the programmer instructing how a computer ought to approach such a task, it is instead given a large dataset, from which it is instructed to learn how to perform the task.

While image generation may be easy for artists, it is a task whose decomposition into a sequence of computer-understandable sub-tasks is more difficult. To approach this task, the architecture of generative adversarial networks [9] proposes to train two neural networks – a generator that produces images, and a critic that returns a score of how well the generator's output resembles the dataset. After training, the generator should be able to produce realistic images. The critic's objective is quite similar to hypothesis testing, for its goal is to discriminate between two datasets, one of actual images and one of generated images, each of which may be considered as a set of samples from an underlying distribution. Therefore, a choice for the critic may be an integral probability metric, which is a metric on the space of probability distributions [1].

[1] Note that in the original GAN formulation, a critic maps from images to the real line, while in the IPM formulation, a critic maps from probability distributions over images to the real line.

In particular, a critic can be implemented as the maximum mean discrepancy, which is a particular integral probability metric that returns the maximum difference that a function in a reproducing kernel Hilbert space achieves between two probability distributions. A modification of this idea was proposed in the paper MMD GAN [5], which used a composition of a Gaussian kernel with an encoder neural network as a critic in the GAN framework. Despite the success of the algorithm, the paper did not attempt to characterize the properties of this encoder kernel after training, which is related to the problem that we will try to address: we will use classical methods from hypothesis testing [14] in order to characterize the behavior of a certain data-dependent kernel that averages the encoder kernel of MMD GAN [5] across a reference distribution. Our contributions are as follows:

1. The methods in [14] are difficult to extend in their original form to high-dimensional data, such as images; we present a generalization of their method by building a data-dependent kernel based on an encoder neural network.

2. Based on the previous contribution, we propose a method for extending the ideas of [14] to high-dimensional data by leveraging the dimensionality reduction capabilities of autoencoder neural networks.

3. We establish that the bound on the difference between the eigenvalues of the integral operators of two positive semi-definite kernels is the same as the bound on the absolute difference of those kernels; i.e. the map from kernels to eigenvalues of their integral operators is continuous in the supremum norm. In doing so, we relate the spectral properties of the kernel of [14] to those of our data-dependent encoder kernel.

In Section 2, we cover some preliminaries on neural networks, reproducing kernel Hilbert spaces, and the background for anisotropic kernels [14]. In Section 3, we relate the integral probability metric induced by the anisotropic kernel to that induced by the encoder kernel. In Section 4, we relate the spectrum of the anisotropic kernel to that of the encoder kernel by proving a more general fact about the relation between the spectra of the integral operators of kernels whose absolute difference is uniformly bounded. In Section 5, we state some asymptotic properties of the encoder kernel and the anisotropic kernel. In Section 6, we outline the permutation test of the IPM of the encoder kernel, and we discuss the extension of [14] to high dimensional data as well as how its method can grant time complexity savings to our encoder kernel. In Section 7, we present experiments validating our ideas.

2 Preliminaries

2.1 Notation

Let $(\Omega, \mathcal{B}(\Omega), \mathbb{P})$ be a probability space with Borel σ-algebra $\mathcal{B}(\Omega)$ [2]. Then $\mathcal{L}^2 = \mathcal{L}^2(\Omega, \mathbb{P})$ is the set of square integrable functions $f : \Omega \mapsto \mathbb{R}$ such that $||f||_{\mathbb{P}}^2 := \int_\Omega f^2 d\mathbb{P} < \infty$. We denote the square summable (resp. summable) sequences on $\mathbb{R}$ as ℓ^2 (resp. ℓ^1); a sequence $a = (a_1, a_2, ...)$ is in ℓ^2 (resp. ℓ^1) if $\sum_{n=1}^\infty a_n^2 < \infty$ (resp. $\sum_{n=1}^\infty |a_n| < \infty$). We denote the set of continuously differentiable functions as $\mathcal{C}^1$; for a $\mathcal{C}^1$ function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$, we denote its Jacobian at x as $\mathbf{J}_f(x)$. Finally, we denote the normal distribution with mean μ and variance σ^2 as $\mathcal{N}(\mu, \sigma^2)$.

[2] While we do not really discuss measure theory, we specify this for precision's sake.

2.2 Neural Networks

Definition 1. *A **feedforward neural network** is a composition of functions*

$$\phi_L \circ \phi_{L-1} \circ ... \circ \phi_1 : \mathbb{R}^{d_1} \mapsto \mathbb{R}^{d_{L+1}},$$

*where for $i = 1, ..., L$, $\phi_i : \mathbb{R}^{d_i} \mapsto \mathbb{R}^{d_{i+1}}$ is defined by $\phi_i(x) = \sigma(\boldsymbol{W}_i x + b_i)$, $\boldsymbol{W}_i \in \mathbb{R}^{d_{i+1} \times d_i}$, $b_i \in \mathbb{R}^{d_{i+1}}$, and $\sigma : \mathbb{R} \mapsto \mathbb{R}$ is a nonlinear function that is applied element-wise. σ is called an **activation function**, and ϕ_i is called a **layer**.*

In the introduction, we also briefly discussed what role an "encoder neural network" plays in MMD GANs. In general, the range of an encoder neural network has much lower dimensionality than its domain does, and when we require certain training objectives (reconstruction error, frequently set to be mean-squared error), heuristically we may say that the encoder network learns a low-dimensional representation of the data, i.e. it is a dimensionality reduction technique. Encoders generally are a composition of layers whose domains gradually decrease in dimensionality, as will be shown in the experiments; conversely, the layers of decoders, whose role is to reconstruct the encoder's input given the encoder's low-dimensional representation, generally increase gradually in dimensionality.

2.3 Reproducing Kernel Hilbert Spaces and Maximum Mean Discrepancy

Definition 2. *Let $\mathcal{X}$ be a (nonempty) set. A symmetric function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is called a **positive semi-definite kernel** if for any $n \in \mathbb{Z}^+$, $c_1, ..., c_n \in \mathbb{R}$, and $x_1, ..., x_n \in \mathcal{X}$, we have $\sum_{i,j=1}^{n,n} c_i c_j k(x_i, x_j) \geq 0$. In other words, the matrix $[k(x_i, x_j)]_{i,j=1,1}^{n,n}$ is positive semi-definite.*

Definition 3. *[7] Let $\mathcal{H}$ be a Hilbert space of functions mapping from some set $\mathcal{X}$ to $\mathbb{R}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. Then $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a **reproducing kernel** of the **reproducing kernel Hilbert space** $\mathcal{H}$ if*

1. *$\forall x \in \mathcal{X}$, $k(\cdot, x) \in \mathcal{H}$.*
2. *$\forall x \in \mathcal{X}$ and $f \in \mathcal{H}$, $f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}$.*

*The second item is known as the **reproducing property**. A function $\phi : \mathcal{X} \mapsto \mathcal{H}$ is called a **feature map** if $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$. More than one such ϕ can exist, but $k(x, \cdot)$ is always a valid feature map.*

A more intuitive description of the functions inhabiting $\mathcal{H}_k$ (the RKHS of a kernel k) is that $f \in \mathcal{H}_k$ if there exist some $x_1, x_2, ... \in \mathcal{X}$ such that $f = \sum_{n=1}^{\infty} a_n k(x_n, \cdot)$ for constants $\{a_n\} \subset \mathbb{R}$.

Definition 4. *Let $\mathbb{P}$ and $\mathbb{Q}$ be two (Borel) probability distributions on $\mathcal{X}$. Given a reproducing kernel Hilbert space $\mathcal{H}$ with kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, [8] defines the **maximum mean discrepancy** between $\mathbb{P}$ and $\mathbb{Q}$ as*

$$\gamma^2 (\mathbb{P}, \mathbb{Q}; k) = \sup_{f \in \mathcal{H}} [\mathbb{E}_{x \sim \mathbb{P}}[f(x)] - \mathbb{E}_{x \sim \mathbb{Q}}[f(x)]]$$

*Additionally, the **mean embedding** of $\mathbb{P}$ with respect to k is $\mu_{\mathbb{P}}^k = \mu_{\mathbb{P}} = \mathbb{E}_{x \sim \mathbb{P}} k(x, \cdot)$.*

Remark 5. *By the reproducing property, we may write the squared MMD as*

$$\gamma^2(\mathbb{P}, \mathbb{Q}; k) = \int_{\mathcal{X}} \int_{\mathcal{X}} k(x, y) d(\mathbb{P} - \mathbb{Q})(x) d(\mathbb{P} - \mathbb{Q})(y) = ||\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}||_{\mathcal{H}}^2$$

by Lemmas 4 and 6 in [8], where $||\cdot||_{\mathcal{H}}$ is the RKHS norm induced by $\langle\cdot,\cdot\rangle_{\mathcal{H}}$. Its squared empirical estimate, given i.i.d. $\{x_i\}_{i=1}^{n_1} \sim \mathbb{P}$ and $\{y_j\}_{j=1}^{n_2} \sim \mathbb{Q}$, is

$$\gamma^2\left(\{x_i\}_{i=1}^{n_1}, \{y_j\}_{j=1}^{n_2}; k\right) = \frac{1}{n_1^2}\sum_{i,j=1}^{n_1,n_1} k(x_i, x_j) - \frac{2}{n_1 n_2}\sum_{i,j=1}^{n_1,n_2} k(x_i, y_j) + \frac{1}{n_2^2}\sum_{i,j=1}^{n_2,n_2} k(y_i, y_j)$$

Definition 6. *[10] A kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is called **characteristic** if for all probability distributions $\mathbb{P}$ on $\mathcal{X}$, the map $\mathbb{P} \mapsto \mathbb{E}_{t\sim\mathbb{P}} k(\cdot, t)$ from the space of Borel probability measures on $\mathcal{X}$ to the RKHS of k is injective. That is, the embedding of a probability distribution is unique. From [8], if a kernel is characteristic, then $\gamma^2(\mathbb{P}, \mathbb{Q}; k) = 0$ if and only if $\mathbb{P} = \mathbb{Q}$.*

2.4 The Gaussian RKHS

In the next section, we will focus on a specific kernel, which is a modification of the Gaussian kernel, a popular choice for use in kernel methods. This section characterizes the RKHS of the Gaussian kernel and summarizes some of the justifications for its use.

Fix $\sigma > 0$ and $d \in \mathbb{Z}^+$. For $f : \mathbb{C}^d \to \mathbb{C}$, define

$$||f||_\sigma = \left(\frac{2^d}{\pi^d \sigma^{2d}} \int_{\mathbb{C}^d} |f(z)|^2 e^{|z-\bar{z}|^2/\sigma^2} dz\right)^{1/2}$$

and

$$\mathcal{H}_\sigma = \{f : \mathbb{C}^d \to \mathbb{C} \text{ s.t. } f \text{ holomorphic and } ||f||_\sigma < \infty\}$$

Theorem 7. *(Theorem 4.38 in [2].) Let $\sigma > 0$ and $d \in \mathbb{Z}^+$. Then $\mathcal{H}_\sigma$ is a RKHS and $k_\sigma : \mathbb{C}^d \mapsto \mathbb{C}$, $k_\sigma(z, z') = \exp(-|z - \bar{z}'|^2/\sigma^2)$, is its reproducing kernel. Furthermore, define for $n \in \mathbb{Z}^+$*

$$e_n(z) = \sqrt{\frac{2^n}{\sigma^{2n} n!}} z^n e^{z^2/\sigma^2} \quad \text{and} \quad \otimes_{j=1}^d e_{n_j}(z_1, ..., z_d) = \prod_{j=1}^d e_{n_j}(z_j)$$

Then $\{\otimes_{j=1}^d e_{n_j} : \mathbb{C}^d \mapsto \mathbb{C}\}_{n_1,...,n_d \in \mathbb{Z}^+}$ is an orthonormal basis for H_σ.

Remark 8. *Note that the elements of this orthonormal basis rapidly decay at infinity and are infinitely differentiable.*

Remark 9. *These results concern the Gaussian RKHS when the domain of the kernel is $\mathbb{C}^d$; they may be made specific to the case where the domain is a subset of $\mathbb{R}^d$. See Chapter 4 of [2] for more details.*

Theorem 10. *(Theorem 4.47 in [2].) Let μ be a finite measure on $\mathbb{R}^d$ or Lebesgue measure, $p \in (1, \infty)$, and $\sigma > 0$. Then the operator $S_{k_\sigma} : \mathcal{L}^p(\mu) \mapsto \mathcal{H}_\sigma(\mathbb{R}^d)$ is injective, where*

$$S_{k_\sigma} g = \int_{\mathbb{R}^d} k_\sigma(\cdot, x) g(x) d\mu(x)$$

Corollary 11. *This implies that the Gaussian kernel is characteristic, and therefore the maximum mean discrepancy induced by the Gaussian kernel is a metric on the space of probability distributions.*

2.5 Background for Anisotropic Kernels

This section follows [14] closely. Let $\mathbb{P}$ and $\mathbb{Q}$ be (Borel) probability measures with support on a compact set $\Omega \subset \mathbb{R}^{d_1}$, with the former measure associated with the null hypothesis and the latter associated with the alternative hypothesis, $0 < \rho_1 < 1$, $\rho_2 := 1 - \rho_1$, and $\mu_R := \rho_1\mathbb{P} + \rho_2\mathbb{Q}$. μ_R is called the reference distribution; while no proof relies on the fact that it is a mixture of the two probability measures, in practice μ_R takes this form, so we may as well specify its form here. ρ_1 and ρ_2 represent the (asymptotic) proportion of datapoints from $\mathbb{P}$ and $\mathbb{Q}$ respectively.

Definition 12. *Let an invertible, symmetric, and positive semi-definite matrix $\Sigma_r \in \mathbb{R}^{d_1 \times d_1}$ be associated with each point in Ω, and define $\{\Sigma_r^{-1}\}_r$ as the* ***covariance field*** [3]*. The* ***asymmetric affinity kernel*** *for some $\sigma > 0$ is defined as*

$$a(r, x) = \exp\left(-\frac{1}{2\sigma^2}(x - r)^T\Sigma_r^{-1}(x - r)\right)$$

While a is not symmetric and therefore not a kernel, we may still define an analogue of its mean embedding of a probability measure $\mathbb{P}$: $\mu_{\mathbb{P}}^a = \mathbb{E}_{x\sim\mathbb{P}}a(\cdot, x)$.

Additionally, we denote $k_{\mathcal{L}^2}(x, y) = \mathbb{E}_{r\sim\mu_R}a(r, x)a(r, y) = \langle a(\cdot, x), a(\cdot, y)\rangle_{\mu_R}$, and we denote the Gaussian kernel by $k_\sigma : \mathbb{R}^{d_2} \times \mathbb{R}^{d_2} \mapsto \mathbb{R}$ with some bandwidth $\sigma > 0$,

$$k_\sigma(x, y) = \exp\left(-\frac{1}{2\sigma^2}|x - y|^2\right).$$

In k_σ, a, and $k_{\mathcal{L}^2}$, the role of σ may be interpreted as controlling how much $|x - y|$ affects how similar each kernel considers some inputs x and y to be, since the same inputs x and y would be considered less similar by a kernel with smaller bandwidth. Compare this to the role of σ^2 in parameterizing the normal distribution.

Remark 13. *The maximum mean discrepancy corresponding to $k_{\mathcal{L}^2}$ is*

$$\gamma^2(\mathbb{P}, \mathbb{Q}; k_{\mathcal{L}^2}) = \int_\Omega |\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}|^2 d\mu_R$$

i.e. the squared $\mathcal{L}^2(\mu_R)$ distance between the mean embeddings of $\mathbb{P}$ and $\mathbb{Q}$.

3 Encoder-Defined Anisotropic Kernels

Definition 14. *Let $f : \mathbb{R}^{d_1} \mapsto \mathbb{R}^{d_2}$. Suppose for all $\varepsilon > 0$, there exists a $\delta > 0$ such that if $|x-y| < \delta$, then $|f(x) - f(y) - \boldsymbol{J}_f(y)(x - y)|/|x - y| < \varepsilon$. Then f is called* ***uniformly differentiable*** [4]*.*

If the domain of f is restricted to the compact set $\Omega \subset \mathbb{R}^{d_1}$ and if for all i and j $\partial f_i/\partial x_j$ is continuous (i.e. f is $\mathcal{C}^1$), then the partial derivatives are also uniformly continuous, which is equivalent to $\mathbf{J}_f$ being uniformly continuous. In the particular case where f is a feedforward neural network with a $\mathcal{C}^1$ activation function, we may conclude that f is uniformly differentiable. In the following theorems, we will show that for any function $f : \mathbb{R}^{d_1} \mapsto \mathbb{R}^{d_2}$ ($d_1 \gg d_2$) that satisfies a slightly stronger form of uniform differentiability and that has a Holder continuous inverse (e.g. if f is an encoder with latent space dimensionality d_2 and if the decoder is Holder continuous), then the kernel $k_\sigma \circ f$

[3] Without loss of generality, the inverse may be replaced by a pseudoinverse.

[4] See Chapter 5 Exercise 8 in [11].

[5]formed by composing the Gaussian kernel $k_\sigma : \mathbb{R}^{d_2} \times \mathbb{R}^{d_2} \mapsto \mathbb{R}$ with f is uniformly close to the asymmetric affinity kernel a with covariance field $\{(\mathbf{J}_f(r)^T\mathbf{J}_f(r))^{-1}\}_r$, provided that both a and $k_\sigma \circ f$ have the same sufficiently small bandwidth.

Theorem 15. *Let $\mathbb{P}$ be a probability distribution, $q \in (0,1]$, and $f : \Omega \subset \mathbb{R}^{d_1} \mapsto \mathbb{R}^{d_2}$ be an injective function satisfying a slightly stronger form of uniform differentiability: $r \in (0,1)$ such that for $\varepsilon \in (0,1)$, there exists δ with $\varepsilon^r \leq \delta^{\frac{2}{q}-1}$ [6] and $|x-r| < \delta \implies |f(x)-f(r)-\boldsymbol{J}_f(r)(x-r)|/|x-r| < \varepsilon$. Suppose $f^{-1} : \text{range}(f) \mapsto \Omega$ is Holder continuous with Holder exponent q [7], and that there exists some $\lambda_{d_2} > 0$ such that $\forall r \in \Omega$, the smallest singular value of $\boldsymbol{J}_f(r)$ is at least λ_{d_2}. Then there exists a bandwidth $\sigma > 0$ such that $\forall r \in \Omega$,*

$$|\mu^a_{\mathbb{P}}(r) - \mu^{k_\sigma \circ f}_{\mathbb{P}}(r)| < \varepsilon,$$

where a has covariance field $\{(\boldsymbol{J}_f^T(r)\boldsymbol{J}_f(r))^{-1}\}_{r\in\Omega}$, is independent of $\mathbb{P}$, and has the same bandwidth as k_σ.

Proof. Firstly, observe that the Holder condition on f^{-1} implies

$$|x - r| = |f^{-1}(f(x)) - f^{-1}(f(r))| \leq \alpha |f(x) - f(r)|^q.$$

Therefore,

$$m(\delta) := \inf_{|x-r|\geq\delta} |f(x) - f(r)| \geq \alpha^{-1/q} \inf_{|x-r|\geq\delta} |x - r|^{1/q} = \alpha^{-1/q}\delta^{1/q}.$$

Let $\varepsilon \in (0,1)$, and suppose $|x - r| \geq \delta$. Then $|\mathbf{J}_f(r)(x - r)| \geq \lambda_{d_2}|x - r| \geq \lambda_{d_2}\delta$ [8], and we have

$$a(r,x) = \exp\left(-\frac{1}{2\sigma^2}|\mathbf{J}_f(r)(x - r)|^2\right) \leq \exp\left(-\frac{1}{2\sigma^2}\lambda^2_{d_2}\delta^2\right)$$

From the above, observe that if we want to set $\sigma > 0$ such that $a(r,x) \leq \varepsilon$, then σ must satisfy $-\frac{1}{2\sigma^2} \leq \lambda_{d_2}^{-2}\delta^{-2}\log(\varepsilon)$. Now consider $k_\sigma \circ f$: since $m(\delta) \geq \alpha^{-1/q}\delta^{1/q}$, we get

$$(k_\sigma \circ f)(r,x) = \exp\left(-\frac{1}{2\sigma^2}|f(x) - f(r)|^2\right) \leq \exp\left(-\frac{1}{2\sigma^2}m(\delta)^2\right) \leq \exp\left(-\frac{1}{2\sigma^2}\alpha^{-2/q}\delta^{2/q}\right).$$

Similarly to the discussion of a, if we want to set σ such that $(k \circ f)(r,x) \leq \varepsilon$, then σ must satisfy $-\frac{1}{2\sigma^2} \leq \alpha^{2/q}\delta^{-2/q}\log(\varepsilon)$. Set $\sigma > 0$ so that

$$-\frac{1}{2\sigma^2} = \delta^{-2/q}\max\{\lambda_{d_2}^{-2}, \alpha^{2/q}\}\log(\varepsilon) \leq \max\{\lambda_{d_2}^{-2}\delta^{-2}, \alpha^{2/q}\delta^{-2/q}\}\log(\varepsilon),$$

when $\delta \leq 1$. By the stronger form of uniform differentiability, there exists a δ such that $\varepsilon^r \leq \delta^{1-\frac{2}{q}} \leq 1$, so that for all $x, r \in \mathbb{R}^{d_1}$ with $|x-r| < \delta$, we have $|f(x)-f(r)-\mathbf{J}_f(r)(x-r)|/|x-r| < \varepsilon$. Now let $|x - r| < \delta$ and L be the Lipschitz constant of $x \mapsto e^{-x^2}$ ($L < 1$). Then

[5]That $k \circ f$ is a kernel is due to the fact that $k(f(x), \cdot)$ is its feature map.

[6]The requirement that $\delta^{\frac{2}{q}-1} \geq \varepsilon^r$ is much more interpretable if $q = 1$ (i.e. if f^{-1} is Lipschitz), since then it merely reads $\delta \geq \varepsilon^r$. Also, the condition could be relaxed by setting a constant $C > 0$ so that $\varepsilon^r \leq C\delta^{\frac{2}{q}-1}$, but we do not do this for sake of simplicity.

[7]That is, there exists some $q \in (0,1]$ and $\alpha > 0$ such that for all $\varepsilon > 0$, there exists $\delta > 0$ such that for all $x, r \in \text{range}(f)$, we have $|f^{-1}(x) - f^{-1}(y)| \leq \alpha|x - y|^q$.

[8]See this StackExchange answer for a review on SVD and lower bounds.

$$
\begin{aligned}
|(k_\sigma \circ f)(r,x) - a(r,x)| &= \left|\exp\left(-\frac{|f(r)-f(x)|^2}{2\sigma^2}\right) - \exp\left(-\frac{|\mathbf{J}_f(r)(x-r)|^2}{2\sigma^2}\right)\right| \\
&\leq \frac{L}{2\sigma^2}\left||f(x)-f(r)| - |\mathbf{J}_f(r)(x-r)|\right| \\
&\leq \frac{L}{2\sigma^2}|f(x)-f(r)-\mathbf{J}_f(r)(x-r)| \\
&< \frac{L\varepsilon\delta}{2\sigma^2} \\
&= L\varepsilon\delta\cdot\delta^{-2/q}\max\{\lambda_{d_2}^{-2},\alpha^{2/q}\}\cdot -\log(\varepsilon) \\
&\leq L\varepsilon\delta\delta^{-\frac{2}{q}}\max\{\lambda_{d_2}^{-2},\alpha^{2/q}\}\cdot -\log(\varepsilon) \\
&\leq L\max\{\lambda_{d_2}^{-2},\alpha^{2/q}\}\varepsilon^{1-r}\cdot -\log(\varepsilon)
\end{aligned}
$$

Therefore with an appropriate setting of σ, on $\{x,r : |x-r| < \delta\}$ the absolute difference between a and $k \circ f$ is bounded uniformly by $\varepsilon^{1-r}\cdot -\log(\varepsilon)$ multiplied by some constant, which approaches 0 as $\varepsilon \to 0^+$, and on $\{x,r : |x-r| \geq \delta\}$ both a and $k \circ f$ are less than ε. Hence

$$
\begin{aligned}
\left|\mu_{\mathbb{P}}^{a}(r) - \mu_{\mathbb{P}}^{k_\sigma \circ f}(r)\right| &= \left|\int a(r,x)d\mathbb{P}(x) - \int (k_\sigma \circ f)(r,x)d\mathbb{P}(x)\right| \\
&\leq \int |a(r,x) - (k_\sigma \circ f)(r,x)|\, d\mathbb{P}(x) \\
&\leq \int \max\left\{L\max\{\lambda_{d_2}^{-2},\alpha^{2/q}\}\varepsilon^{1-r}\cdot -\log(\varepsilon),\varepsilon\right\}d\mathbb{P}(x) \\
&= \max\{L\max\{\lambda_{d_2}^{-2},\alpha^{2/q}\}\varepsilon^{1-r}\cdot -\log(\varepsilon),\varepsilon\}
\end{aligned}
$$

□

Remark 16. *We briefly comment on the requirement that there exists a $\lambda_{d_2} > 0$ such that $\forall r \in \Omega$, the smallest singular value of $\boldsymbol{J}_f(r)$ is bounded below by λ_{d_2}: if the smallest singular value of some $\boldsymbol{J}_f(r)$ were 0, then there is a direction local to r that $\boldsymbol{J}_f(r)$ finds "useless" and therefore along which there is insufficient decay when $|x-r| \geq \delta$. This requirement may be replaced if one can tolerate multiplying a by a mollified (continuous) indicator function for the set $\{x,r : |x-r| < \delta\}$ so that sufficient decay is achieved.*

The hypothesis in the above theorem that f is injective becomes clearer when f is an encoder neural network, for the requirement of injectivity may be interpreted as barring two different datapoints from having the same low-dimensional representation. The hypothesis that f^{-1} is Holder implies that f separates $\{|x-r| < \delta\}$ from $\{|x-r| \geq \delta\}$ quickly enough.

What the above theorem shows is that when the bandwidths of a and $k_\sigma \circ f$ are small enough, both kernels are more sensitive to the distance their inputs have in the ambient space Ω; in particular, under the assumptions of the previous theorem (particularly the lower bound on the smallest singular value), when $|x-r|$ are even somewhat far apart and when σ is sufficiently large, the behavior of both kernels is almost entirely determined by the local information provided by f at r, i.e. $\mathbf{J}_f(r)$. This suggests that equipping $k_{\mathcal{L}^2}$ with the covariance field $\{(\mathbf{J}_f^T(r)\mathbf{J}_f(r))^{-1}\}_r$ is a valid choice, if one wants to take advantage of the data pre-processing capabilities of f. Additionally, note that the bound between $k_\sigma \circ f$ and a when $|x-r| < \delta$ is made worse by smaller values for σ^2, since the only

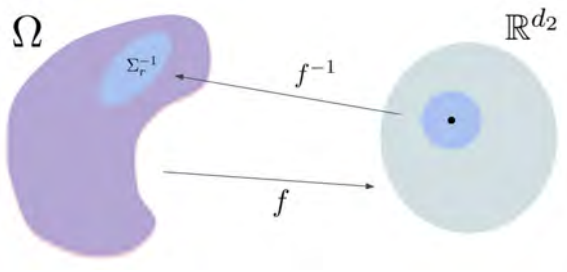


Figure 1: f maps from the ambient space Ω to the latent space $\mathbb{R}^{d_2}$. $k_\sigma \circ f$ first pre-processes data via f and then applies a standard Gaussian kernel, while a simply computes a local linear transformation based on $\mathbf{J}_f$. However, since the preimage of a small symmetric ball in $\mathbb{R}^{d_2}$ is pulled back to the covariance matrix $\Sigma_r^{-1} = (\mathbf{J}_f^T(r)\mathbf{J}_f(r))^{-1}$ in Ω, $a(r,x) \approx (k_\sigma \circ f)(r,x)$.

role of σ^2 is to force decay when $|x-r| \geq \delta$; this implies that setting σ^2 to be much smaller than what ε forces it to be will substantially degrade the upper bound, and the behavior of these kernels may diverge. The theorem also motivates the definition of a new kernel:

Definition 17. *Let $k_\sigma : \mathbb{R}^{d_2} \times \mathbb{R}^{d_2} \mapsto \mathbb{R}$ denote the Gaussian kernel with bandwidth σ, and let $f : \Omega \mapsto \mathbb{R}^{d_2}$ be an encoder. Denote $\phi(x) = (k_\sigma \circ f)(\cdot, x)$, and define the* ***encoder kernel*** *as*

$$k_f(x,y) = \langle \phi(x), \phi(y) \rangle_{\mu_R} = \int_\Omega \phi(x)\phi(y) d\mu_R$$

so that ϕ is a feature map for k_f.

Remark 18. *As with the MMD of $k_{\mathcal{L}^2}$, the maximum mean discrepancy of k_f between probability distributions $\mathbb{P}$ and $\mathbb{Q}$ may be written as the $\mathcal{L}^2(\mu_R)$ inner product between their mean embeddings by ϕ, i.e.*

$$\gamma^2(\mathbb{P}, \mathbb{Q}; k_f) = \int_\Omega |\mu_\mathbb{P} - \mu_\mathbb{Q}|^2 d\mu_R = ||\mu_\mathbb{P}^\phi - \mu_\mathbb{Q}^\phi||_{\mu_R}^2.$$

Because the encoder f is designed to be trained to discover clearer, low-dimensional latent representations of data from Ω, an intuitive description of the role of $\phi = k_\sigma \circ f$ is that f first finds the latent representations of inputs r and x, which is then followed by the Gaussian kernel computing the similarity between their filtered representations. Following $k_{\mathcal{L}^2}$'s tactic of taking advantage of the $\mathcal{L}^2(\mu_R)$ inner product to employ a reference set and build a data-dependent kernel, we then integrate $\phi(x)\phi(y)$ against $r \sim \mu_R$ to obtain k_f, to which the above theorem shows that $k_{\mathcal{L}^2}$ is an approximation when its covariance field is $\{(\mathbf{J}_f^T(r)\mathbf{J}_f(r))^{-1}\}_r$.

To finish this section, we provide a short proof that if two feature maps are close, then their maximum mean discrepancies are also close; in our situation, the feature maps are a and ϕ.

Theorem 19. *Let $\mathbb{P}$ and $\mathbb{Q}$ be probability distributions, $\Omega \subset \mathbb{R}^{d_1}$ be compact, $f : \Omega \mapsto \mathbb{R}^{d_2}$ be a uniformly differentiable and injective function, $\varepsilon > 0$, and*

$$\left|\mu_\mathbb{P}^a(r) - \mu_\mathbb{P}^\phi(x)\right| < \varepsilon, \qquad \left|\mu_\mathbb{Q}^a(r) - \mu_\mathbb{Q}^\phi(r)\right| < \varepsilon$$

for all $r \in \Omega$. In particular, this requirement is satisfied when the asymmetric affinity kernel a has the covariance field $\{(\boldsymbol{J}_f^T(r)\boldsymbol{J}_f(r))^{-1}\}_{r\in\Omega}$ and both ϕ and a have the same bandwidth as specified in the previous theorem. Then

$$\left|\gamma^2(\mathbb{P},\mathbb{Q};k_f)-\gamma^2(\mathbb{P},\mathbb{Q};k_{\mathcal{L}^2})\right|<6\varepsilon.$$

Proof. By the definition of $\gamma^2(\mathbb{P},\mathbb{Q};k_{\mathcal{L}^2})$ and $\gamma^2(\mathbb{P},\mathbb{Q};k_f)$, we have

$$\begin{aligned}\left|\gamma^2(\mathbb{P},\mathbb{Q};k_{\mathcal{L}^2})-\gamma^2(\mathbb{P},\mathbb{Q};k_f)\right| &= \left|\int|\mu^a_{\mathbb{P}}-\mu^a_{\mathbb{Q}}|^2d\mu_R-\int|\mu^\phi_{\mathbb{P}}-\mu^\phi_{\mathbb{Q}}|^2d\mu_R\right| \\ &\leq \int\left|(\mu^a_{\mathbb{P}})^2-(\mu^\phi_{\mathbb{P}})^2\right|+\left|\left(\mu^a_{\mathbb{Q}}\right)^2-(\mu^\phi_{\mathbb{Q}})^2\right|+2\left|\mu^\phi_{\mathbb{P}}\mu^\phi_{\mathbb{Q}}-\mu^a_{\mathbb{P}}\mu^a_{\mathbb{Q}}\right|d\mu_R \\ &< 4\varepsilon+\int\left|\mu^\phi_{\mathbb{P}}\left(\mu^\phi_{\mathbb{Q}}-\mu^a_{\mathbb{Q}}\right)+\mu^a_{\mathbb{Q}}\left(\mu^\phi_{\mathbb{P}}-\mu^a_{\mathbb{P}}\right)\right|d\mu_R \\ &< 6\varepsilon.\end{aligned}$$

□

4 Spectral Analysis

It may not be entirely clear how $k_{\mathcal{L}^2}$ and k_f are related to each other, especially in terms of their spectra. In this section, we prove a general fact about the eigenvalues of kernel integral operators and apply it to the spectra of $k_{\mathcal{L}^2}$ and k_f. Weyl's inequality shows that the eigenvalues of a Hermitian matrix are stable under perturbation, where the norm of the perturbation may be taken to be the spectral norm of matrices. Since the Gram matrix of a kernel is Hermitian, one should expect this fact to extend to kernel integral operators, and in fact this is what we prove in the following theorem.

Theorem 20. *Let k,l: $\mathbb{R}^d\times\mathbb{R}^d\mapsto\mathbb{R}$ be two bounded positive semi-definite kernels, $\mathbb{P}$ a probability measure, and suppose there exists some $\eta>0$ such that for all x and $y\in\mathbb{R}^d$, $|k(x,y)-l(x,y)|<\eta$. Let $S_k:\mathcal{L}^2(\mathbb{R}^d,\mathbb{P})\mapsto\mathcal{H}_k$ and $S_l:\mathcal{L}^2(\mathbb{R}^d,\mathbb{P})\mapsto\mathcal{H}_l$ be the integral operators of k and l, where $\mathcal{H}_k$ is the RKHS of k, $S_kg=\int k(x,\cdot)g(x)d\mathbb{P}(x)$, and S_l and $\mathcal{H}_l$ are similarly defined. If λ_i and ν_i are the i-th greatest eigenvalues of S_k and S_l respectively (which exist by Mercer's theorem [2]) with respect to $\mathbb{P}$, then $|\lambda_i-\nu_i|<\eta$.*

Remark 21. *In other words, for all $i\in\mathbb{Z}^+$, if $\mathcal{K}$ is the convex cone of bounded, positive semi-definite kernels and if $\lambda_i(k)$ is the i-th greatest eigenvalue of the integral operator of k, then $k\mapsto\lambda_i(k):\mathcal{K}\mapsto\mathbb{R}^+$ is continuous with respect to the supremum norm.*

Proof. Let $n\in\mathbb{Z}^+$, δ_{ij} be the Kronecker delta and $\{x_i\}_{i=1}^n\subset\mathbb{R}^d$. Define the Gram matrices with zero diagonal $\mathbf{K}_n=\frac{1}{n}[(1-\delta_{ij})k(x_i,x_j)]_{i,j=1}^n$ and $\mathbf{L}_n=\frac{1}{n}[(1-\delta_{ij})l(x_i,x_j)]_{i,j=1}^n$, and define the perturbations $\varepsilon_{ij}=k(x_i,x_j)-l(x_i,x_j)$ that by assumption satisfy $|\varepsilon_{ij}|\leq\eta$ for all $i,j=1,...,n$. Define the hollow perturbation matrix $\mathbf{R}_n=\frac{1}{n}[(1-\delta_{ij})\varepsilon_{ij}]_{i,j=1}^n$, so that $\mathbf{K}_n-\mathbf{L}_n=\mathbf{R}_n$. Then for all $v\in\mathbb{R}^n$ with $|v|=1$,

$$|\mathbf{R}_nv|^2<\sum_i\left(\sum_j\frac{\varepsilon_{ij}}{n}v_j\right)^2\leq\sum_i\left(\sum_j\frac{\varepsilon_{ij}^2}{n^2}\right)\left(\sum_jv_j^2\right)=\frac{1}{n^2}\sum_{ij}\varepsilon_{ij}^2=\frac{1}{n^2}\sum_{ij}\eta^2=\eta^2,$$

where the strict inequality comes from the zeroed diagonal of $\mathbf{R}_n$, and the following inequality comes from Cauchy-Schwartz. Therefore the spectral norm of $\mathbf{R}_n$ is (strictly) less than η.

Denote the spectrum of a linear operator T as $\lambda(T)$, so the spectrum of $\mathbf{K}_n$ may be written in decreasing order as $\lambda(\mathbf{K}_n) = \{\bar{\lambda}_1, ..., \bar{\lambda}_n\}$ and those of $\mathbf{L}_n$ as $\lambda(\mathbf{L}_n) = \{\bar{\nu}_1, ..., \bar{\nu}_n\}$. (The spectra may be written in decreasing order since $\mathbf{K}_n$ and $\mathbf{L}_n$ are Hermitian, which is a consequence of their entries being determined by (symmetric) kernels.) By a more specific form of Weyl's inequality (Equation 13 in [12]), we have that for $i, j = 1, ..., n$, $|\bar{\lambda}_i - \bar{\nu}_i| < \eta$.

Definition 22. *[13] For x, $y \in \ell^2(\mathbb{R})$, their **rearrangement distance** is*

$$\delta_2(x, y) = \left[\inf_{\sigma} \sum_{i=1}^{\infty} (x_i - y_{\sigma(i)})^2\right]^{1/2},$$

where σ is a permutation over $\mathbb{Z}^+$.

Let $x, x' \sim \mathbb{P}$. By the boundedness of k and l, we have that $\mathbb{E}_{x,x'} k^2(x, x') < \infty$ and $\mathbb{E}_{x,x'} l^2(x, x') < \infty$. Therefore by Theorem 3.1 in [13], $\delta_2(\lambda(\mathbf{K}_n), \lambda(S_k)) \to 0$ and $\delta_2(\lambda(\mathbf{L}_n), \lambda(S_l)) \to 0$ as $n \to \infty$.

The following lemma was said to be "clear" on page 117 in [13], but for the sake of completeness, we include its proof here.

Lemma 23. *Let $a = (a_1, a_2, ...)$ and $b = (b_1, b_2, ...)$ be sequences in $\ell^2(\mathbb{R})$, such that $a_1 \geq a_2 \geq ... \geq 0$ and $b_1 \geq b_2 \geq ... \geq 0$. Then $\delta_2^2(a, b) = \sum_{i=1}^{\infty} (a_i - b_i)^2$.*

Proof. Firstly, let $a_1 \geq a_2 > 0$, $b_1 \geq b_2 > 0$. Then $a_1 b_1 + a_2 b_2 = a_1(b_1 - b_2) + b_2(a_2 + a_1) \geq a_2(b_1 - b_2) + b_2(a_2 + a_1) = a_2 b_1 + a_1 b_2$. But this implies

$$a_1^2 + a_2^2 + b_1^2 + b_2^2 - 2a_1 b_1 - 2a_2 b_2 \leq a_1^2 + a_2^2 + b_1^2 + b_2^2 - 2a_2 b_1 - 2a_1 b_2.$$

Thus $(a_1 - b_1)^2 + (a_2 - b_2)^2 = \text{LHS} \leq \text{RHS} = (a_2 - b_1)^2 + (a_1 - b_2)^2$. For the proof of the lemma, note that since a permutation is essentially an infinite number of swaps of the kind in the assumption, for any permutation σ we have

$$\sum_{i=1}^{\infty} (a_i - b_i)^2 \leq \sum_{i=1}^{\infty} (a_i - b_{\sigma(i)})^2,$$

and therefore the LHS is $\delta_2^2(a, b)$. □

By Theorem A.5.13 in [2], the eigenvalues of the integral operators S_k and S_l, which are nonnegative by the positive semi-definiteness of k and l, may be sorted in descending order as $\lambda(S_k) = \{\lambda_1, \lambda_2, ...\}$ and $\lambda(S_l) = \{\nu_1, \nu_2, ...\}$. By the lemma, we get that $\delta_2^2(\lambda(\mathbf{K}_n), \lambda(S_k)) = \sum_{i=1}^{\infty} (\lambda_i - \bar{\lambda}_i)^2 \to 0$ and $\delta_2^2(\lambda(\mathbf{L}_n), \lambda(S_l)) = \sum_{i=1}^{\infty} (\nu_i - \bar{\nu}_i)^2 \to 0$ (having padded $\lambda(\mathbf{K}_n)$ and $\lambda(\mathbf{L}_n)$ with infinitely many 0's, as convention dictates). Fix some $i \in \mathbb{Z}^+$ and $\varepsilon > 0$. Now there exists some $N \in \mathbb{Z}^+$, $N > i$, such that $n > N$ implies that both $\delta_2^2(\lambda(\mathbf{K}_n), \lambda(S_k))$ and $\delta_2^2(\lambda(\mathbf{L}_n), \lambda(S_l))$ are less than $\varepsilon/2$. Then we have the following bound:

$$\begin{aligned}
|\lambda_i - \nu_i| &\leq |\lambda_i - \bar{\lambda}_i| + |\bar{\lambda}_i - \bar{\nu}_i| + |\bar{\nu}_i - \nu_i| \\
&\leq \delta_2(\lambda(\mathbf{K}_n), \lambda(S_k)) + \eta + \delta_2(\lambda(\mathbf{L}_n), \lambda(S_l)) \\
&< \frac{\varepsilon}{2} + \eta + \frac{\varepsilon}{2} \\
&= \eta + \varepsilon,
\end{aligned}$$

where the second inequality comes from the fact that $|\lambda_i - \bar{\lambda}_i| \leq \sqrt{\sum_{j=1}^{\infty}(\lambda_j - \bar{\lambda}_j)^2}$ (likewise for ν_i and $\bar{\nu}_i$) and the earlier discussion using Weyl's inequality. Since $|\lambda_i - \nu_i| < \eta + \varepsilon$ for arbitrary ε, the theorem follows. □

Remark 24. *We should note that by Mercer's theorem (Theorem 4.49 in [2]), $\lambda_i \to 0$ and $\nu_i \to 0$, so the above bound on the difference between eigenvalues of S_k and S_l is nontrivial for at least finitely many eigenvalues when η is not too large.*

During the proof bounding the difference between the mean embeddings of a and ϕ, we proved that $|a(\cdot, x) - \phi(x)|$ is uniformly bounded (say by some $\eta > 0$). Therefore, $|k_{\mathcal{L}^2}(x, y) - k_f(x, y)|$ is uniformly bounded by 2η, and the above theorem shows that the individual differences between their eigenvalues is also bounded by 2η.

5 Analysis of Testing Power

5.1 Facts About the Kernel

Suppose $k(x, y) = \int \varphi(s, x)\varphi(s, y)d\mu(s)$, where μ is a probability measure and for all $x \in \Omega$, $\varphi(\cdot, x) \in \mathcal{L}^2(\Omega, \mu)$ is continuous. Then k is also continuous. Since $k_{\mathcal{L}^2}$ and k_f are the $\mathcal{L}^2(\mu_R)$ inner products of $a(\cdot, x)$ with $a(\cdot, y)$ and $\phi(x)$ with $\phi(y)$ respectively, both of which are continuous and square integrable, we see that both kernels are continuous. $k \in \{k_{\mathcal{L}^2}, k_f\}$ is positive semi-definite and continuous, so by Mercer's theorem, we have an expansion

$$k(x, y) = \sum_k \lambda_k \psi_k(x)\psi_k(y),$$

that converges absolutely and uniformly, and where $\{\psi_k\}$ is an orthonormal set of functions with respect to $\mu_R := \rho_1\mathbb{P} + \rho_2\mathbb{Q}$. We assume that when $\mathbb{P} \neq \mathbb{Q}$, there exists $l \in \mathbb{Z}^+$ where $\lambda_l > 0$ and

$$\int \psi_l(x)d(\mathbb{P} - \mathbb{Q})(x) \neq 0.$$

Roughly speaking, because Mercer's theorem shows that $\{\psi_k\}$ is an orthonormal basis for $\mathcal{L}^2(\Omega, \mathbb{P})$, we may interpret this assumption as saying that $\mathbb{P}$ and $\mathbb{Q}$ differ in a direction spanned by some ψ_k.

5.2 The Centered Kernel

From now on, we denote the maximum difference between the MMD's as η (which was established in Section 3); more explicitly,

$$\left|\gamma^2(\mathbb{P}, \mathbb{Q}; k_f) - \gamma^2(\mathbb{P}, \mathbb{Q}; k_{\mathcal{L}^2})\right| < \eta$$

provided that the feature map a of $k_{\mathcal{L}^2}$ has the form specified in Section 3 (covariance field given by $\{(\mathbf{J}_f^T(r)\mathbf{J}_f(r))^{-1}\}_r$ and with small enough σ). With $\{x_i\}_{i=1}^{n_1}$ and $\{y_j\}_{j=1}^{n_2}$ as *i.i.d.* samples from $\mathbb{P}$ and $\mathbb{Q}$ respectively, for convenience denote

$$\gamma_n^2(k_{\mathcal{L}^2}) := \gamma^2\left(\{x_i\}_{i=1}^{n_1}, \{y_j\}_{j=1}^{n_2}; k_{\mathcal{L}^2}\right) \quad \text{and} \quad \gamma_n^2(k_f) := \gamma^2\left(\{x_i\}_{i=1}^{n_1}, \{y_j\}_{j=1}^{n_2}; k_f\right)$$

to be the empirical MMDs of $k_{\mathcal{L}^2}$ and k_f between $\{x_i\}$ and $\{y_j\}$. Note that the previous bound η also holds for MMDs between empirical distributions, so having finished the approximation of

$\gamma^2(\mathbb{P}, \mathbb{Q}; k_f)$ with $\gamma^2(\mathbb{P}, \mathbb{Q}; k_{\mathcal{L}^2})$, we may start relating the behavior of $\gamma_n^2(k_f)$ to that of $\gamma_n^2(k_{\mathcal{L}^2})$, using the results of [14].

Definition 25. *Let $\mathbb{P}$ be the probability measure associated with the null hypothesis. With $k = k_{\mathcal{L}^2}$ or $k = k_f$, define the **centered kernel** $\tilde{k}$ as*

$$\tilde{k}(x, y) = k(x, y) - \mu_{\mathbb{P}}(x) - \mu_{\mathbb{P}}(y) + \mathbb{E}_{(s,t)\sim\mathbb{P}\otimes\mathbb{P}} k(s, t)$$

Remark 26. *It may be shown that $\gamma^2(\mathbb{P}, \mathbb{Q}; k) = \gamma^2(\mathbb{P}, \mathbb{Q}; \tilde{k})$ and so $\gamma_n^2(k) = \gamma_n^2(k_f)$.*

Proposition 27. *$\tilde{k} \in \{\tilde{k}_{\mathcal{L}^2}, \tilde{k}_f\}$ is positive semi-definite on $(\Omega, \mathbb{P})$. Additionally, if $k \in \{k_{\mathcal{L}^2}, k_f\}$ is continuous, then the following hold:*

1) $\tilde{k}$ is continuous, and $\forall x, y \in \Omega, 0 \leq \tilde{k}(x, x) \leq 4$ and $|\tilde{k}(x, y)| \leq 4$.

2) $\tilde{k}$ has the spectral expansion

$$\tilde{k}(x, y) = \sum_k \tilde{\lambda}_k \tilde{\psi}_k(x) \tilde{\psi}_k(y)$$

where $\tilde{\lambda}_k > 0$ for all k, $\mathbb{E}_{\mathbb{P}} \tilde{\psi}_k = 0$, $\{\tilde{\psi}_k\}$ is an orthonormal set of continuous functions on $\mathcal{L}^2(\Omega, \mathbb{P})$, $\{\tilde{\lambda}_k\} \in \ell^1 \cap \ell^2$, and the expansion converges absolutely and uniformly.

3) The eigenfunctions $\{\tilde{\psi}_k\}$ are (square) integrable with respect to the alternative hypothesis $\mathbb{Q}$. Furthermore, $\sum_k \tilde{\lambda} \mathbb{E}_{\mathbb{Q}} \tilde{\psi}_k^2 \leq 4$.

Proof. We make the simple observation that this is Proposition 3.3 in [14], whose proof makes *no use* of the particular form of $k_{\mathcal{L}^2}(x, y) = \int a(r, x) a(r, y) d\mu_R(r)$, and so their proof works for k_f as well. □

5.3 Limiting Distribution of MMD and Asymptotic Consistency

Let us place some restrictions on $\mathbb{P}$ and $\mathbb{Q}$: we assume that the probability density function of $\mathbb{Q}$ is $q_\tau = p + \tau g$, where p is the density of $\mathbb{P}$, g is some function such that q_τ is a density, and $\tau \in [0, 1]$. Define $c_k = \mathbb{E}_g \tilde{\psi}_k$, where $\tilde{\psi}_k$ is an eigenfunction of the integral operator of $\tilde{k}$ with respect to $\mathbb{P}$.

Theorem 28. *Suppose $k \in \{k_{\mathcal{L}^2}, k_f\}$, $n_1, n_2 \to \infty$ such that $n_1/n \to \rho_1$, $n_2/n \to \rho_2 := 1 - \rho_1$, and $n := n_1 + n_2$.*

1) If $\frac{\tau}{n^{-1/2}} \to a$ for $a \in [0, \infty)$ (with τ potentially depending on n), then

$$n \cdot \gamma_n^2(k) \xrightarrow{d} \sum_k \tilde{\lambda}_k (-a c_k + \xi_k)^2,$$

where $\xi_k \sim \mathcal{N}\left(0, \frac{1}{\rho_1} + \frac{1}{\rho_2}\right)$.

2) If $\tau = 1$, then with $T = \sum_k \tilde{\lambda}_k c_k^2$,

$$\sqrt{n}\left(\gamma_n^2(k) - T\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

where $\sigma^2 = 4\left(\frac{1}{\rho_1} \sum_k \tilde{\lambda}_k c_k^2 + \frac{1}{\rho_2} \sum_{kl} \tilde{\lambda}_k \tilde{\lambda}_l c_k c_l S_{kl}\right)$, $S_{kl} = \mathbb{E}_q \tilde{\psi}_k \tilde{\psi}_l - c_k c_l$.

Proof. We make the simple observation that the proof of Theorem 3.4 in [14] makes *no use* of the fact that $k_{\mathcal{L}^2}(x, y) = \int a(r, x) a(r, y) d\mu_R(r)$. Therefore, the theorem works for k_f as well. □

Corollary 29. *From (1), if $\tau = 0$, then*

$$n \cdot \gamma_n^2(k) \xrightarrow{d} \sum_k \tilde{\lambda}_k \xi_k^2$$

Corollary 30. *Suppose $|k_{\mathcal{L}^2} - k_f| < \eta$. Then the limiting distributions of $\gamma_n^2(k_{\mathcal{L}^2})$ and $\gamma_n^2(k_f)$ when $\tau = 1$ differ in mean by at most η.*

Proof. Using Mercer's theorem, we can get the expansions

$$\tilde{k}_{\mathcal{L}^2}(x,y) = \sum_k \tilde{\lambda}_k \tilde{\psi}_k(x)\tilde{\psi}_k(y) \qquad \tilde{k}_f(x,y) = \sum_k \tilde{\nu}_k \tilde{\varphi}_k(x)\tilde{\varphi}_k(y)$$

so that we may define $c_k = \mathbb{E}_g \tilde{\psi}_k$, $d_k = \mathbb{E}_g \tilde{\varphi}_k$, $T_1 = \sum_k \tilde{\lambda}_k c_k^2$, and $T_2 = \sum_k \tilde{\nu}_k d_k^2$. Then

$$\begin{aligned}
\sum_k \tilde{\lambda}_k c_k^2 &= \sum_k \tilde{\lambda}_k \left(\int \tilde{\psi}_k(x) g(x) dx \right) \left(\int \tilde{\psi}_k(y) g(y) dy \right) \\
&= \sum_k \tilde{\lambda}_k \int \int \tilde{\psi}_k(x)\tilde{\psi}_k(y) g(x) g(y) dx dy \\
&= \int \int \left[\sum_k \tilde{\lambda}_k \tilde{\psi}_k(x)\tilde{\psi}_k(y) \right] g(x) g(y) dx dy \\
&= \int \int \tilde{k}_{\mathcal{L}^2}(x,y) g(x) g(y) dx dy
\end{aligned}$$

where the first line is by definition, the second line and third lines are by Fubini's theorem, and the fourth is by the expansion given by Mercer's theorem. Similarly $T_2 = \int \int \tilde{k}_f(x,y) g(x) g(y) dx dy$. Therefore, if we apply the previous theorem's second part to get

$$\sqrt{n}\left(\gamma_n^2(k_{\mathcal{L}^2}) - T_1\right) \xrightarrow{d} \mathcal{N}(0, \sigma_1^2) \quad \text{and} \quad \sqrt{n}\left(\gamma_n^2(k_f) - T_2\right) \xrightarrow{d} \mathcal{N}(0, \sigma_2^2),$$

where σ_1 and σ_2 are determined by the formula for σ in part 2, then the corollary follows. □

We now relate the cumulative distribution functions of $\gamma_n^2(k_{\mathcal{L}^2})$ and $\gamma_n^2(k_f)$ using the previous bound η on their absolute difference and simple arguments.

Corollary 31. *Denote $Z = \sum_k \tilde{\lambda}_k (-ac_k + \xi_k)^2$, the limiting distribution of $\gamma_n^2(k_{\mathcal{L}^2})$ under the null hypothesis. With the assumptions of Theorem 28 and $|\gamma_n^2(k_{\mathcal{L}^2}) - \gamma_n^2(k_f)| < \eta$, and with F denoting the cumulative distribution function, we have for $t \in \mathbb{R}$:*

$$F_Z(t) < \liminf_{n \to \infty} F_{\gamma_n^2(k_f)}\left(\frac{t}{n} + \eta\right).$$

Proof. Firstly, Theorem 12 from [8] proves the existence of the limiting random variable (in the sense of convergence in distribution) to which $\gamma_n^2(k_f)$ converges as $n \to \infty$. For $n \in \mathbb{Z}^+$, since $|\gamma_n^2(k_{\mathcal{L}^2}) - \gamma_n^2(k_f)| < \eta$ and the MMD is nonnegative, we have that $\gamma_n^2(k_{\mathcal{L}^2}) \in \left[0, \frac{t}{n}\right]$ implies $\gamma_n^2(k_f) \in \left[0, \frac{t}{n} + \eta\right]$, and therefore

$$F_{\gamma_n^2(k_{\mathcal{L}^2})}\left(\frac{t}{n}\right) \leq F_{\gamma_n^2(k_f)}\left(\frac{t}{n} + \eta\right).$$

But from Theorem 29, we know

$$\lim_{n\to\infty} F_{\gamma_n^2(k_{\mathcal{L}^2})}\left(\frac{t}{n}\right) = F_Z(t)$$

and therefore, if we fix $\varepsilon > 0$, then there exists N with $n \geq N \implies |F_Z(t) - F_{\gamma_n^2(k_{\mathcal{L}^2})}(t)| < \varepsilon$. Combining this with the previous statements, we get

$$F_Z(t) - \varepsilon < F_{\gamma_n^2(k_{\mathcal{L}^2})}\left(\frac{t}{n}\right) \leq F_{\gamma_n^2(k_f)}\left(\frac{t}{n} + \eta\right)$$

implying

$$F_Z(t) < \liminf_{n\to\infty} F_{\gamma_n^2(k_f)}\left(\frac{t}{n} + \eta\right).$$

□

6 Permutation Tests

Here we present an algorithm for the use of the maximum mean discrepancy of $k \in \{k_{\mathcal{L}^2}, k_f\}$ in a permutation test. The first algorithm computes the empirical MMD between two datasets $X = \{x_k\}_{k=1}^{n_1} \sim \mathbb{P}$ and $Y = \{y_l\}_{l=1}^{n_2} \sim \mathbb{Q}$; the second algorithm is a standard permutation test. Of course, a significance level $\alpha > 0$ must be set before the test; if the empirical p-value returned by the second algorithm is less than α, then we reject the null hypothesis (i.e. $\mathbb{P} \neq \mathbb{Q}$); otherwise, we fail to reject the null. To compute k, we must first select a reference set $R = \{r_j\}_{j=1}^{n_R}$, which is a subset of the data. The algorithms here are modified from [14].

6.1 High-Dimensional Covariance Field Selection

[14] proposed a different method to select the covariance field of $k_{\mathcal{L}^2}$, which was by local PCA, i.e. pick a neighborhood for each datapoint and perform PCA only on datapoints within that neighborhood. However, this is infeasible for high-dimensional data because 1) the time complexity of PCA is $O(d^2n + d^3)$ (where n is the number of datapoints and d is the number of features), although admittedly the SVD method can yield time complexity of $O(ndk)$ where k is the number of the selected principal components [9] 2) more importantly, in high dimensions, local PCA requires much more data in order to be accurate. Rather, we propose to leverage the dimensionality reduction capabilities of autoencoders by instead selecting as the covariance field $\{(\mathbf{J}_f^T(r)\mathbf{J}_f(r))^{-1}\}_r$ for some encoder f at every reference point r. (Of course, this is only sensible when there is reason to suspect that low-dimensional structure is hiding in the high-dimensional data.) This yields an approximation to the kernel k_f, which may have superior performance due to its theoretical capability of incorporating nonlinear information during its dimensionality reduction.

6.2 Time Complexity Improvements on $\gamma_n^2(k_f)$

Notice that by the construction of k_f, each evaluation of k_f requires many matrix multiplications, which is exacerbated by the fact that modern neural networks frequently involve tens or hundreds of matrix multiplications per evaluation of each input due to their depth. On the other hand,

[9] Link to a lecture on the time complexity of PCA

when $k_{\mathcal{L}^2}$ is equipped with the covariance field $\{(\mathbf{J}_f^T(r)\mathbf{J}_f(r))^{-1}\}_r$, it can still take advantage of the dimensionality reduction abilities of the encoder yet only require one matrix-vector multiplication per evaluation on a pair of datapoints.

Algorithm 1 Compute the empirical MMD between $X = \{x_k\}_{k=1}^{n_1}$ and $Y = \{y_l\}_{l=1}^{n_2}$ with reference set $R = \{r_j\}_{j=1}^{n_R}$ using the kernel $k_{\mathcal{L}^2}$ or k_f. If $k_{\mathcal{L}^2}$, then $\varphi = a$; if k_f, then $\varphi = \phi$.

function EMPIRICALMMD(X, Y, R, φ)
  $\varphi(R, X) \leftarrow [\varphi(r_j, x_k)]_{j,k=1,1}^{n_R,n_1}$ ▷ evaluate the image of R and X under φ
  $\varphi(R, Y) \leftarrow [\varphi(r_j, y_l)]_{j,l=1,1}^{n_R,n_2}$ ▷ evaluate the image of R and Y under φ
  $\bar{\mu}_X \leftarrow \left[\frac{1}{n_1}\sum_{k=1}^{n_1}\varphi(R, X)_{jk}\right]_{j=1}^{n_R}$ ▷ evaluate the mean embedding of X on R
  $\bar{\mu}_Y \leftarrow \left[\frac{1}{n_2}\sum_{l=1}^{n_2}\varphi(R, Y)_{jl}\right]_{j=1}^{n_R}$ ▷ evaluate the mean embedding of Y on R
  return $\frac{1}{n_R}\sum_{j=1}^{n_R}\left((\bar{\mu}_X)_j - (\bar{\mu}_Y)_j\right)^2$ ▷ compute the $\mathcal{L}^2$ distance between $\bar{\mu}_X$ and $\bar{\mu}_Y$
end function

Algorithm 2 Permutation test for $k_{\mathcal{L}^2}$ or k_f (φ is as in Algorithm 1); $n_{perms} \in \mathbb{Z}^+$ is the number of permutations. Returns the empirical p-value.

$z \leftarrow$ EMPIRICALMMD($\{x_k\}_{k=1}^{n_1}, \{y_l\}_{l=1}^{n_2}, R, \varphi$)
$\vec{m} \leftarrow \vec{0} \in \mathbb{R}^{n_{perms}}$
for $n = 1, 2, ..., n_{perms}$ **do**
  $\{z_i\}_{i=1}^{n_1+n_2} \leftarrow \text{shuffle}\left(\{x_k\}_{k=1}^{n_1} \cup \{y_l\}_{l=1}^{n_2}\right)$
  $\vec{m}_n \leftarrow$ EMPIRICALMMD$\left(\{z_i\}_{i=1}^{n_1}, \{z_i\}_{i=n_1+1}^{n_1+n_2}, R, \varphi\right)$
end for
return $\frac{1}{n_{perms}}\sum_{l=1}^{n_{perms}}\mathbb{1}[\vec{m}_l \geq z]$ ▷ i.e. the average number of times $\vec{m}_l > z$

7 Experiments

On the dataset MNIST [15], a database of 28 x 28 grayscale images of 70,000 handwritten digits, we train a convolutional autoencoder with encoder f and decoder h: $f = f_\theta : \mathbb{R}^{28\times28} \mapsto \mathbb{R}^6$ has parameters θ, and $h = h_\phi : \mathbb{R}^6 \mapsto \mathbb{R}^{28\times28}$ has parameters ϕ. They are trained to to minimize mean squared loss:

$$\inf_{\theta,\phi} \mathbb{E}_{x\sim\nu}|x - h(f(x))|^2$$

where ν is the distribution generating all MNIST images [10]. Training was run over 50 epochs, with 0.001 initial learning rate, and using the Adam optimizer [6]. The dataset was split into 60,000 training examples and 10,000 test examples. The details of f and h are described below; while the results of the experiment do not seem to depend significantly on the choices of activation functions, we pick $\mathcal{C}^1$ functions for consistency.

Define the exponential linear unit activation function ELU [11]:

[10] The loss is 0 if $h = f^{-1}$, which would satisfy the requirement of injectivity of f in the proof bounding the difference between the mean embedding of a and $k \circ f$.

[11] ELU documentation

$$\text{ELU}(x) = \begin{cases} x & x > 0 \\ e^x - 1 & x \leq 0 \end{cases}$$

Let $\textsc{Conv}_1$ denote a 2D convolutional layer [12] producing 8 channels from 1, $\textsc{Conv}_2$ produce 16 from 8, and $\textsc{Conv}_3$ produce 32 from 16. Let $\mathbf{W}_1$ be a matrix mapping from $\mathbb{R}^{3\times3\times32}$ to $\mathbb{R}^{128}$, and let $\mathbf{W}_2$ be a matrix mapping from $\mathbb{R}^{128}$ to $\mathbb{R}^6$. Then the architecture of f is

$$f = \mathbf{W}_2 \circ \text{ELU} \circ \mathbf{W}_1 \circ \text{ELU} \circ \textsc{Conv}_3 \circ \text{ELU} \circ \textsc{Conv}_2 \circ \text{ELU} \circ \textsc{Conv}_1.$$

Let $\textsc{Conv}_3^T$ denote a 2D transposed convolutional layer [13] producing 16 channels from 32, $\textsc{Conv}_2^T$ produce 8 from 16, and $\textsc{Conv}_1^T$ produce 1 from 8. Let $\mathbf{W}_2$ be a matrix mapping $\mathbb{R}^6$ to $\mathbb{R}^{128}$, and let $\mathbf{W}_1$ be a matrix mapping $\mathbb{R}^{128}$ to $\mathbb{R}^{3\times3\times32}$. Let $\sigma(z) : \mathbb{R} \mapsto \mathbb{R}$ denote the sigmoid function. Then the architecture of h is

$$h = \sigma \circ \textsc{Conv}_1^T \circ \text{ELU} \circ \textsc{Conv}_2^T \circ \text{ELU} \circ \textsc{Conv}_3^T \circ \mathbf{W}_1 \circ \text{ELU} \circ \mathbf{W}_2$$

As usual, the activation functions ELU and σ are applied elementwise. In f and h, each $\textsc{Conv}_i$ and $\textsc{Conv}_i^T$ $(i = 1, 2, 3)$ has kernel side-length 3, stride 2, and padding 1. All code is available on GitHub and was written in PyTorch [1] and NumPy [4]. Other than the choice of activation function, many of the details of f and h were sourced from [3].

For all experiments, we set $\alpha = 0.05$ as usual. $k_{\mathcal{L}^2}$ has the covariance field $\{(\mathbf{J}_f^T(r)\mathbf{J}_f(r))^{-1}\}_{r\in R}$, and k_f remains as before. We compare the performance of $\gamma_n^2(k_{\mathcal{L}^2})$ and $\gamma_n^2(k_f)$ to $\gamma_n^2(k_\sigma)$, where $k_\sigma : \mathbb{R}^{28\times28} \times \mathbb{R}^{28\times28} \mapsto \mathbb{R}$ is the original Gaussian kernel. Permutation tests with k_σ are done with the same procedure as in Algorithm 2 but with EmpiricalMMD calling the squared empirical estimate of MMD, not Algorithm 1. We tested bandwidths $\sigma \in \{2^i\}_{i=-3}^3$ and present the results of the test where $\sigma^2 = 2^1$. Graphs of other bandwidths are included in the appendix, including the bandwidth $\sigma_{auto}^2 = 28^{-2}$ [14] for the Gaussian kernel. Both $k_{\mathcal{L}^2}$ and k_f generally outperform k_σ.

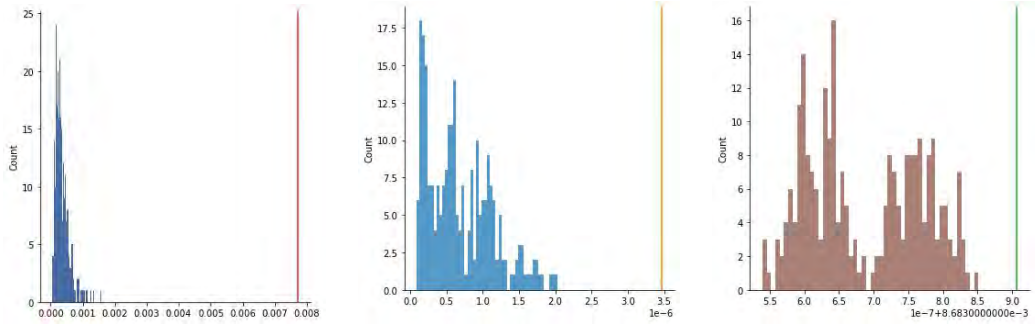
7.1 $\mathbb{P} \neq \mathbb{Q}$

Let $\mathbb{P}$ to be the probability distribution of 8's in MNIST, from which we draw the dataset X of 224 images, and $\mathbb{Q}$ to be the distribution of 3's, from which we draw the dataset Y of 237 images. We chose $\mathbb{P}$ and $\mathbb{Q}$ based on how difficult the encoder found separating each class's latent representations. To be more specific, let S_n $(n = 0, 1, ..., 9)$ represent the set of images from MNIST with label n, $f(S_n) \subset \mathbb{R}^6$ be the image of S_n under f, and

$$\bar{x}_n = |S_n|^{-1} \left[\sum_{x \in S_n} x_j \right]_{j=1}^6.$$

We picked 8 and 3 to compare because $|\bar{x}_8 - \bar{x}_3| = \min_{n\neq m} |\bar{x}_n - \bar{x}_m|$. That is, among all pairs of (different) digits, on average f mapped images of 8's and images of 3's the closest in the latent space $\mathbb{R}^6$. In hindsight, this pair is the obvious choice since 8's and 3's are visually somewhat similar. We plot the histograms produced by the permutation tests, which were ran with 250 permutations. $k_{\mathcal{L}^2}$ and k_f are estimated using a reference set R, which is a mixture of 51 images sampled from $\mathbb{P}$ and $\mathbb{Q}$.

[12] Conv2d documentation
[13] ConvTranspose2d documentation
[14] See the "gamma" keyword argument in sklearn's support vector classifier.

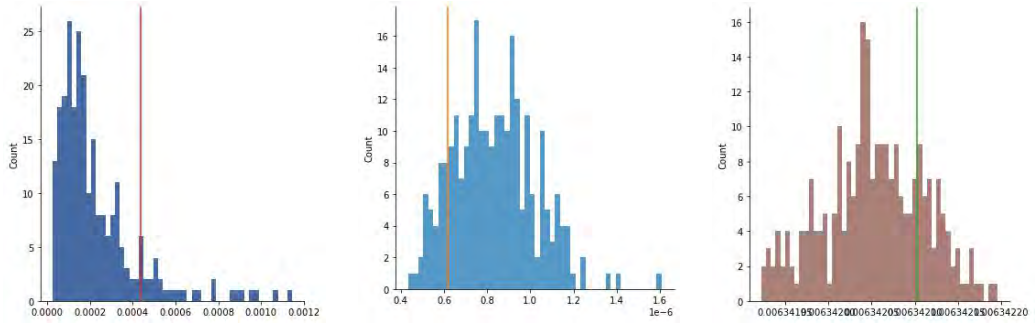


(a) $\gamma_n^2(k_{\mathcal{L}^2})$; p=0.0; time: 16s. (b) $\gamma_n^2(k_f)$; p=0.0; time: 28s. (c) $\gamma_n^2(k_\sigma)$; p=0.0; time: 3s.

Figure 2: Histograms from permutation tests when $\mathbb{P} \neq \mathbb{Q}$. For each kernel, we list the empirical p-value p and the time needed to run the test. The location of p on each graph is the vertical line.

7.2 $\mathbb{P} = \mathbb{Q}$

Set $\mathbb{P}$ to be as before, from which we draw the datasets X_1 of 347 images and X_2 of 289 images. We chose $\mathbb{P}$ to be the distribution on which we test the performance of the kernels under the null hypothesis because 8 has the highest mean-squared reconstruction error. We plot the histograms produced by the permutation tests, which were ran with 250 permutations. $k_{\mathcal{L}^2}$ and k_f are estimated using the reference set R, which is a set of 144 images also sampled from $\mathbb{P}$.



(a) $\gamma_n^2(k_{\mathcal{L}^2})$; p=0.120; time: 44s. (b) $\gamma_n^2(k_f)$; p=0.868; time: 59s. (c) $\gamma_n^2(k_\sigma)$; p=0.224; time: 5s

Figure 3: Histograms from permutation tests when $\mathbb{P} = \mathbb{Q}$.

7.3 Discussion of Experiments

Even after setting random seeds, the permutation test results are not deterministic. However, whether the tests based on k_f or $k_{\mathcal{L}^2}$ reject or fail to reject the null hypothesis rarely changes; code for verifying this is in our GitHub repository. Also, even though in theory f finds 8's and 3's to be the most difficult to distinguish, the experiments based on distinguishing 5's from 8's suggest that the truth is more complicated; in those, $k_{\mathcal{L}^2}$ sometimes outperforms k_f. Finally, since k_f generally produced distributions with extremely pronounced peaks when $\sigma^2 \leq 1$, it seems that k_f is more sensitive to bandwidth selection than does $k_{\mathcal{L}^2}$, whose performance did not depend greatly on σ^2; however, we did observe that for $\sigma^2 \geq 1$, k_f had consistently good performance.

Contrasting the performance of $k_{\mathcal{L}^2}$ when $\mathbb{P} \neq \mathbb{Q}$, it appears that $k_{\mathcal{L}^2}$ is particularly sensitive to differences in the datasets being compared, even when they are drawn from the same distribution. Illustrating this is our experience running experiments for the case where $\mathbb{P} = \mathbb{Q}$, in which we previously ran experiments with X_1 having 299 images, X_2 having 289 images, and R having 67 images. In those experiments, the empirical p-values of $k_{\mathcal{L}^2}$ fluctuated across the decision boundary. This suggests that in this experiment, the sample sizes, including the size of R, are slightly more than the number of samples that $k_{\mathcal{L}^2}$ needs to reliably reject or fail to reject the null hypothesis. Last but not least, we note that the performance of k_σ with $\sigma^2 = \sigma^2_{auto}$ does not separate the datasets as much as $k_{\mathcal{L}^2}$ does when $\mathbb{P} = \mathbb{Q}$, but at the cost of worse separation when $\mathbb{P} \neq \mathbb{Q}$.

8 Conclusion

We proposed a new hypothesis test for high dimensional data using a data-dependent encoder kernel k_f. We extended the concepts of [14] to high dimensional data by proposing a method for choosing their covariance field: taking the "squared" Jacobian of a trained encoder f at each reference point. We extend the analysis of the theoretical behavior of $\gamma^2_n(k_{\mathcal{L}^2})$ to the behavior of $\gamma^2_n(k_f)$, and we also relate their behaviors via bounds on the difference of $\gamma^2_n(k_f)$ and $\gamma^2_n(k_{\mathcal{L}^2})$. Additionally, we proved how the spectrum of $k_{\mathcal{L}^2}$ is related to that of k_f. Finally, rather than using potentially expensive evaluations of a pre-trained encoder neural network f to conduct an accurate hypothesis test with k_f, we can use $k_{\mathcal{L}^2}$ to conduct a hypothesis test with similar accuracy.

9 Acknowledgements

I would like to thank my advisor, Professor Alex Cloninger, for mentoring me on this project. Thank you for patiently explaining difficult concepts, giving valuable feedback on my ideas, proposing new directions, and overall making this project an enjoyable experience.

References

[1] Soumith Chintala Gregory Chanan Adam Paszke, Sam Gross and Zeming Lin Alban Desmaison Luca Antiga Adam Lerer Edward Yang, Zachary DeVito. Automatic differentiation in pytorch. 2017.

[2] Ingo Steinwart Andreas Christmann. *Support Vector Machines*. Springer New York, NY, 2008.

[3] Eugenia Anello. Convolutional autoencoder in pytorch on mnist dataset, 2021.

[4] et al. Charles R. Harris, K. Jarrod Millman. Array programming with NumPy. *Nature*, 585(7825):357–362, sep 2020.

[5] Yu Cheng Yiming Yang Barnabas Pozcos Chun-Liang Li, Wei-cheng Chang. Mmd gan: towards deeper understanding of moment matching network. In *NIPS 2017*, pages 2200–2210. ACM Press, 2017.

[6] Jimmy Ba Diederik P. Kingma. Adam: A method for stochastic optimization. 2015.

[7] Arthur Gretton. Introduction to rkhs, and some simple kernel algorithms, 2019.

[8] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

[9] Mehdi Mirza Bing Xu David Warde-Farley Sherjil Ozair Aaron Courville Yoshua Bengio Ian Goodfellow, Jean Pouget-Abadie. Generative adversarial networks. *Communications of the ACM*, 63:139–144, 2020.

[10] X. Sun B. Scholkopf K. Fukumizu, A. Gretton. Kernel measures of conditional dependence. *Advances in Neural Information Processing Systems*, pages 489–496, 2008.

[11] Walter Rudin. *Principles of Mathematical Analysis*. McGraw Hill, 1967.

[12] Terence Tao. 254a, notes 3a: Eigenvalues and sums of hermitian matrices, 2010.

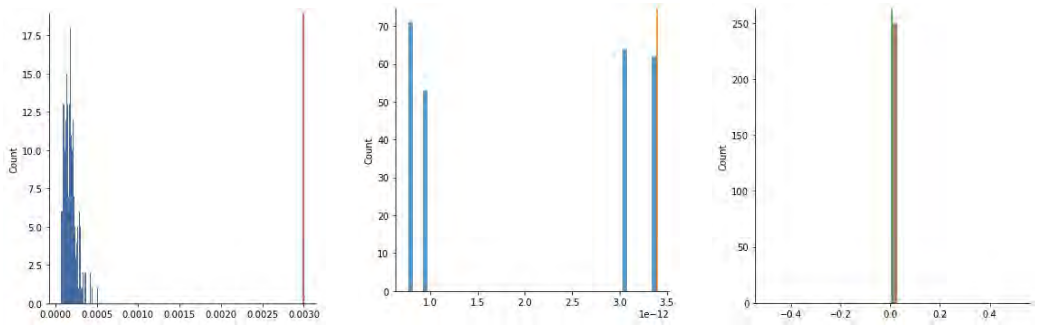
[13] Evarist Gine Vladimir Koltchinskii. Random matrix approximation of spectra of integral operators. *Bernoulli Journal*, 6:113–167, 2000.

[14] Ronald C. Coifman Xiuyuan Cheng, Alexander Cloninger. Two-sample statistics based on anisotropic kernels. *Information and Inference*, pages 677–719, 2020.

[15] Y. Bengio P. Haffner Y. Lecun, L. Bottou. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 11:2278 – 2324, 1998.

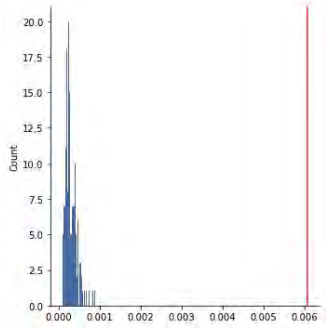
10 Appendix

10.1 $\mathbb{P} \neq \mathbb{Q}$

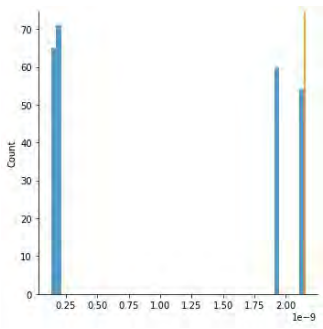


(a) $\gamma_n^2(k_{\mathcal{L}^2})$; p=0.0; time: 16s. (b) $\gamma_n^2(k_f)$; p=0.0; time: 26s. (c) $\gamma_n^2(k_\sigma)$; p=0.0; time: 3s.

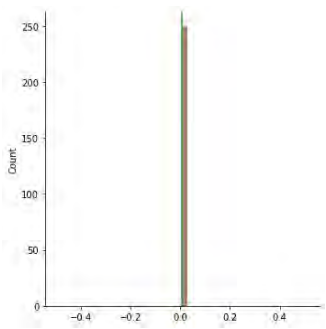
Figure 4: $\sigma^2 = 2^{-3}$



(a) $\gamma_n^2(k_{\mathcal{L}^2})$; p=0.0; time: 19s.

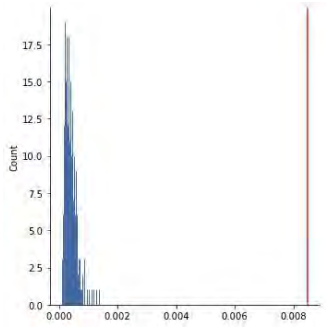


(b) $\gamma_n^2(k_f)$; p=0.0; time: 29s.

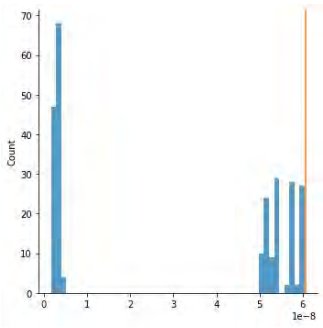


(c) $\gamma_n^2(k_\sigma)$; p=0.0; time: 3s.

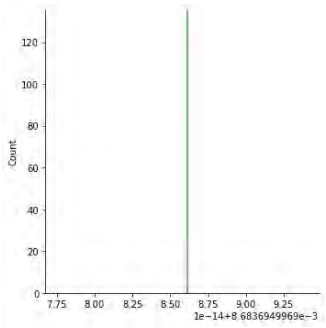
Figure 5: $\sigma^2 = 2^{-2}$



(a) $\gamma_n^2(k_{\mathcal{L}^2})$; p=0.0; time: 16s.

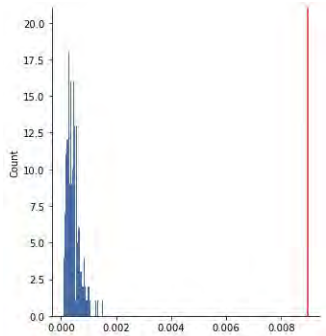


(b) $\gamma_n^2(k_f)$; p=0.0; time: 29s.

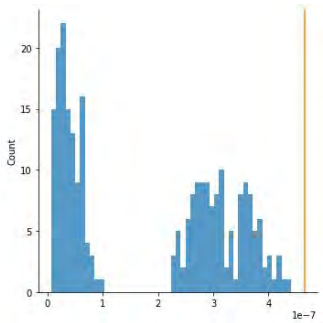


(c) $\gamma_n^2(k_\sigma)$; p=0.0; time: 4s.

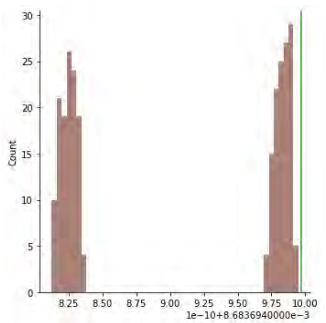
Figure 6: $\sigma^2 = 2^{-1}$



(a) $\gamma_n^2(k_{\mathcal{L}^2})$; p=0.0; time: 16s.

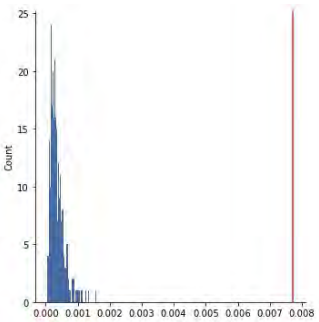


(b) $\gamma_n^2(k_f)$; p=0.0; time: 28s.

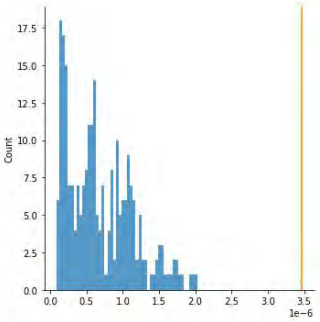


(c) $\gamma_n^2(k_\sigma)$; p=0.0; time: 3s.

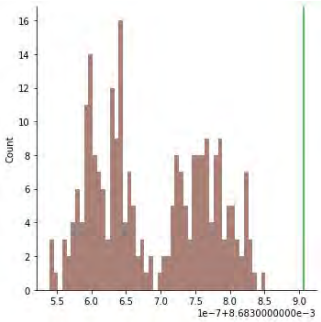
Figure 7: $\sigma^2 = 2^0$



(a) $\gamma_n^2(k_{\mathcal{L}^2})$; p=0.0; time: 16s.

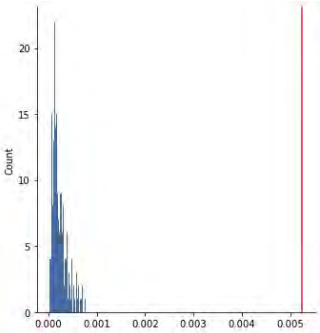


(b) $\gamma_n^2(k_f)$; p=0.0; time: 28s.

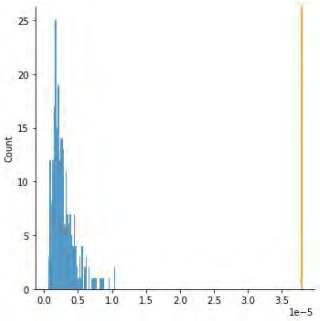


(c) $\gamma_n^2(k_\sigma)$; p=0.0; time: 3s.

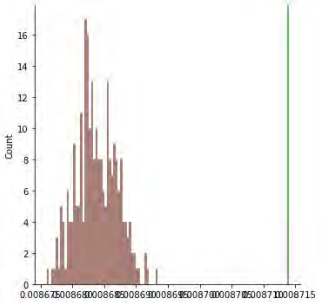
Figure 8: $\sigma^2 = 2^1$



(a) $\gamma_n^2(k_{\mathcal{L}^2})$; p=0.0; time: 16s.

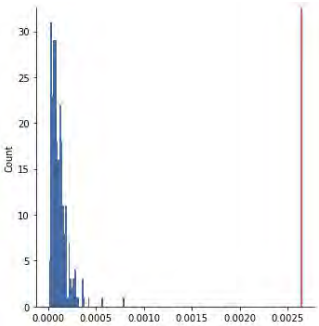


(b) $\gamma_n^2(k_f)$; p=0.0; time: 29s.

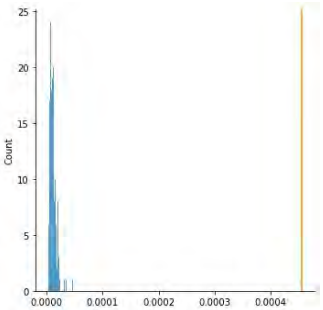


(c) $\gamma_n^2(k_\sigma)$; p=0.0; time: 4s.

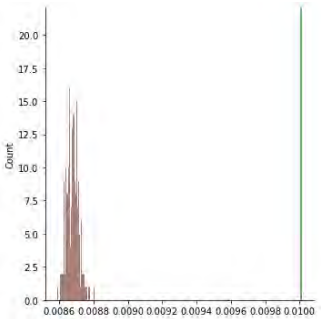
Figure 9: $\sigma^2 = 2^2$



(a) $\gamma_n^2(k_{\mathcal{L}^2})$; p=0.0; time: 15s.



(b) $\gamma_n^2(k_f)$; p=0.0; time: 29s.



(c) $\gamma_n^2(k_\sigma)$; p=0.0; time: 4s.

Figure 10: $\sigma^2 = 2^3$

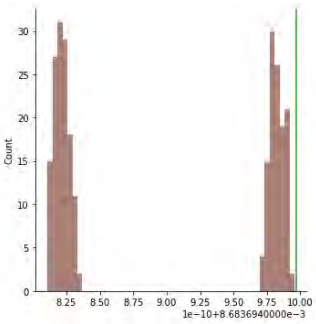
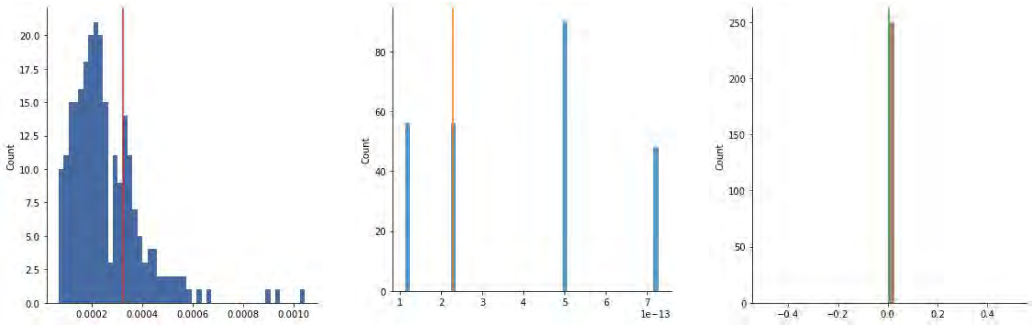


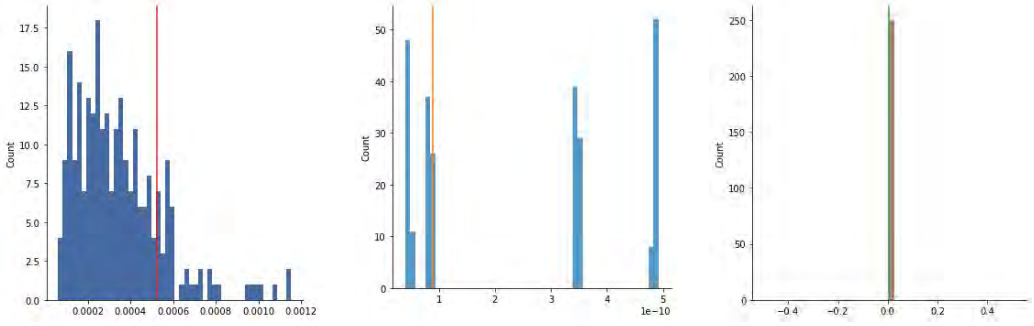
Figure 11: $\gamma_n^2(k_\sigma)$; p=0.0; time: 4s; $\sigma^2 = \sigma^2_{auto}$

10.2 $\mathbb{P} = \mathbb{Q}$



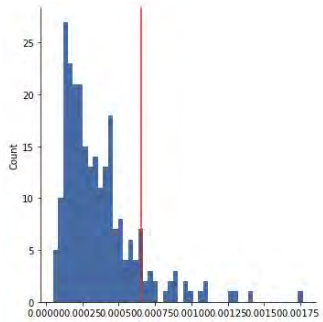
(a) $\gamma_n^2(k_{\mathcal{L}^2})$; p=0.26; time: 43s. (b) $\gamma_n^2(k_f)$; p=0.564; time: 42s. (c) $\gamma_n^2(k_\sigma)$; p=0.0; time: 5s.

Figure 12: $\sigma^2 = 2^{-3}$

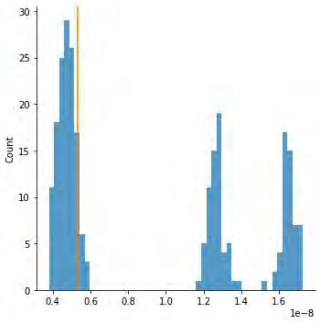


(a) $\gamma_n^2(k_{\mathcal{L}^2})$; p=0.168; time: 44s. (b) $\gamma_n^2(k_f)$; p=0.528; time: 55s. (c) $\gamma_n^2(k_\sigma)$; p=0.0; time: 5s.

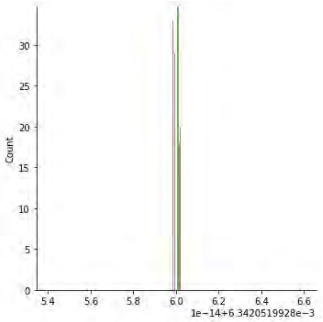
Figure 13: $\sigma^2 = 2^{-2}$



(a) $\gamma_n^2(k_{\mathcal{L}^2})$; p=0.104; time: 44s.

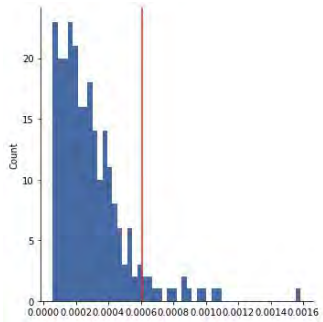


(b) $\gamma_n^2(k_f)$; p=0.512; time: 55s.

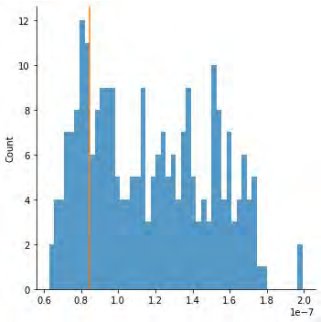


(c) $\gamma_n^2(k_\sigma)$; p=0.300; time: 5s.

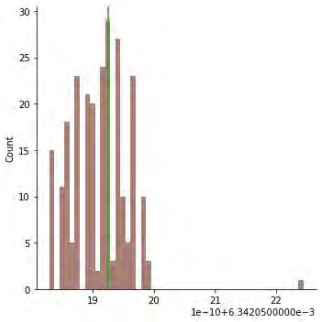
Figure 14: $\sigma^2 = 2^{-1}$



(a) $\gamma_n^2(k_{\mathcal{L}^2})$; p=0.064; time: 41s.

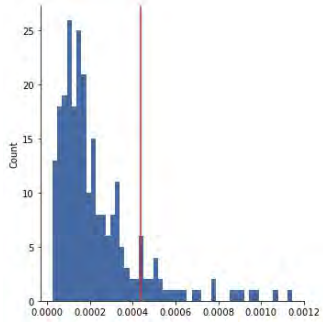


(b) $\gamma_n^2(k_f)$; p=0.776; time: 56s.

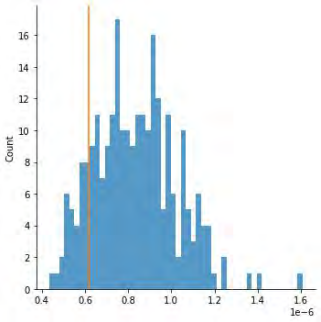


(c) $\gamma_n^2(k_\sigma)$; p=0.352; time: 5s.

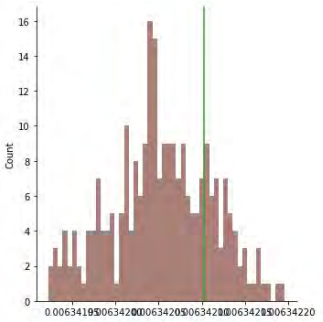
Figure 15: $\sigma^2 = 2^0$



(a) $\gamma_n^2(k_{\mathcal{L}^2})$; p=0.120; time: 44s.

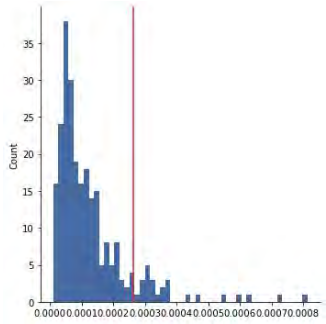


(b) $\gamma_n^2(k_f)$; p=0.868; time: 59s.

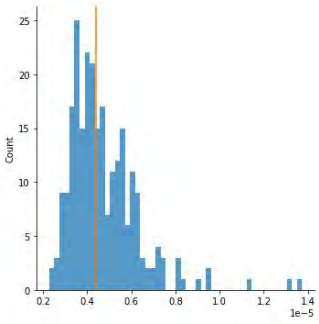


(c) $\gamma_n^2(k_\sigma)$; p=0.224; time: 5s.

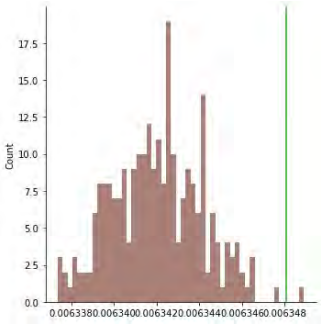
Figure 16: $\sigma^2 = 2^1$



(a) $\gamma_n^2(k_{\mathcal{L}^2})$; p=0.108; time: 43s.

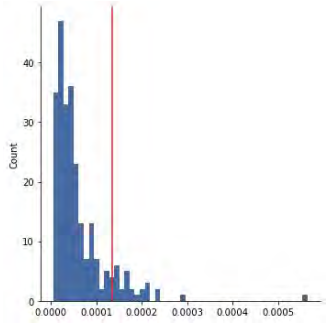


(b) $\gamma_n^2(k_f)$; p=0.492; time: 59s.

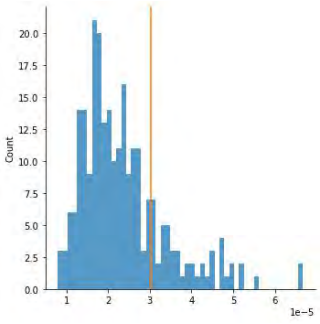


(c) $\gamma_n^2(k_\sigma)$; p=0.004; time: 5s.

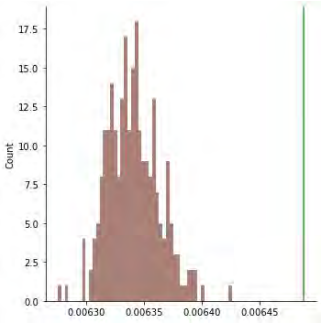
Figure 17: $\sigma^2 = 2^2$



(a) $\gamma_n^2(k_{\mathcal{L}^2})$; p=0.100; time: 41s.



(b) $\gamma_n^2(k_f)$; p=0.196; time: 56s.



(c) $\gamma_n^2(k_\sigma)$; p=0.0; time: 5s.

Figure 18: $\sigma^2 = 2^3$

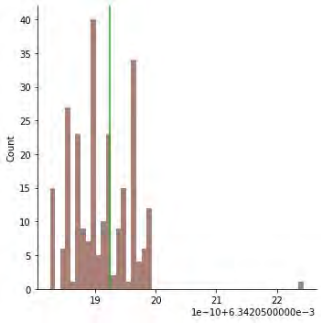


Figure 19: $\gamma_n^2(k_\sigma)$; p=0.356; time: 5s; $\sigma^2 = \sigma^2_{auto}$