A Discrete RKHS Standpoint for Nyström MMD

Farah Cherfaoui, Hachem Kadri, Sandrine Anthoine, Liva Ralaivola

A Discrete RKHS Standpoint for Nyström MMD

Farah Cherfaoui[1,2], Hachem Kadri[1], Sandrine Anthoine[2], et Liva Ralaivola[3]

[1]Aix Marseille Univ, CNRS, LIS, Marseille, France
[2]Aix Marseille Univ, CNRS, I2M, Marseille, France
[3]Criteo AI Lab, Paris, France

Résumé

Maximum mean discrepancy (MMD) is a kernel-based distance measure between probability distributions. It relies on the concept of mean embedding of distributions in a Reproducing Kernel Hilbert Space (RKHS). In this work, we describe a new link between probability distributions and kernel methods. We build upon recent and elegant results on RKHSs over discrete domains which possess novel and appealing properties compared to their continuous counterparts. Based on the observation that discrete RKHSs can contain the Dirac masses, we propose a novel framework for representing and comparing probability distributions. We show how MMD and its fast approximation, Nyström MMD, can be retrieved from the discrete RKHS framework. Our results provide an explanation why MMD and Nyström MMD with a large class of kernels, including graph kernels, remains effective in practice. Our approach is empirically illustrated in the context of three-sample testing.

Key words : MMD kernel mean embedding discrete RKHS Nyström approximation.

1 Introduction

Since they were firstly proposed to learn nonlinear decisions functions with Support Vector Machines [BGV92], kernel methods have enjoyed continuous scientific interest. These methods exploit training data through implicit definition of a similarity between data points that can be expressed as a dot product in a reproducing kernel Hilbert space (RKHS) and have become very popular in many fields [HSS08]. They are acknowledged to have a strong theoretical basis, to be powerful tools for generalizing linear statistical approaches to nonlinear settings, and to be effective in handling structured data [SS02, STC04]. The notion of kernels as dot products in Hilbert spaces was first brought to machine learning by Aizerman et al. [ABR64], while the theoretical foundation of reproducing kernels and their Hilbert spaces dates back to at least Aronszajn [Aro50]. A recent success in this field is *kernel mean embedding* of probability distributions, a framework for representing probabilities in RKHSs [MFSS17]. This makes it possible to use the power of kernel methods to deal with probabilistic modeling and statistical inference problems. Most of the literature to date involves continuous RKHSs, i.e., RKHSs over continuous domains. Jorgensen and Tian [JT15] recently introduced and studied *discrete RKHSs*, spaces that possess novel appealing properties compared to their continuous counterparts : they for example can contain Dirac measures while continuous RKHSs cannot. From this point of view, the main motivation of this work is to shed light on reproducing kernel Hilbert spaces over discrete sets and their role in machine learning.

The authors of [JT15] gave a characterization of RKHSs defined on a countable infinite discrete set V which contain the Dirac masses δ_x for all points $x \in V$. This is a remarkable result, as it offers new alternatives for representing probability distributions in RKHSs. In this paper, we follow this approach and investigate the relation between probability distributions and kernel methods. Specifically, we make the following contributions :

- we provide a discrete RKHS-based framework for characterizing and comparing probability distributions,
- we show how the probability distance measure called maximum mean discrepancy (MMD) can be retrieved from this framework,
- we propose a well-founded fast approximation of MMD based on the Nyström method (Nyström MMD) that justifies the use of Nyström approxi-

mation with MMD,

- we show the effectiveness of Nyström MMD in the context of three-sample testing.

2 Background

In this section we start by giving some background about kernel mean embedding of probability distributions and maximum mean discrepancy (MMD). We then review the basics of discrete RKHSs.

2.1 Kernel mean embedding and MMD

Kernel mean embedding provides an RKHS-based interface between kernel methods and probability distributions. For a textbook reference, we refer the reader to [MFSS17]. Let $\mathcal{H}$ be an RKHS of functions on a separable topological space $\mathcal{X}$ with a continuous and bounded kernel k, i.e. $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $\sup_{x\in\mathcal{X}} k(x,x) < \infty$. The kernel mean embedding $\mu_{\mathbb{P}}$ of a distribution $\mathbb{P}$ is $\mu_{\mathbb{P}} := \int_{\mathcal{X}} k(x,\cdot)\,\mathrm{d}\mathbb{P}(x)$, and given samples $\{x_i\}_{i=1}^n$ from $\mathbb{P}$, $\hat{\mu}_{\mathbb{P}} := \frac{1}{n}\sum_{i=1}^n k(x_i,\cdot)$ is an empirical estimate of $\mu_{\mathbb{P}}$. The kernel mean embedding helps define a metric for probability distributions, the *maximum mean discrepancy* [GBR+12a], which, for two probability distributions $\mathbb{P}$ and $\mathbb{Q}$, is the RKHS distance between their mean embeddings :

$$\mathrm{MMD}_k(\mathbb{P},\mathbb{Q}) := \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}.$$

A fundamental concept underlying kernel mean embedding is the notion of *characteristic* kernel. A kernel k is said to be characteristic if the map $\mu : \mathbb{P} \to \mu_{\mathbb{P}}$ is injective. This is crucial because it ensures that $\mathrm{MMD}_k(\mathbb{P},\mathbb{Q}) = 0$ if and only if $\mathbb{P} = \mathbb{Q}$. Given i.i.d. samples $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^m$ from $\mathbb{P}$ and $\mathbb{Q}$, respectively, an empirical estimate of $\mathrm{MMD}_k(\mathbb{P},\mathbb{Q})$ can be obtained as $\mathrm{MMD}_k(\hat{\mathbb{P}},\hat{\mathbb{Q}}) = \|\hat{\mu}_{\mathbb{P}} - \hat{\mu}_{\mathbb{Q}}\|_{\mathcal{H}}$, where $\hat{\mathbb{P}}$ and $\hat{\mathbb{Q}}$ are the empirical distributions corresponding to $\mathbb{P}$ and $\mathbb{Q}$, respectively. Using the kernel trick [ABR64], it is easy to see that

$$\mathrm{MMD}_k^2(\hat{\mathbb{P}},\hat{\mathbb{Q}}) = \frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i,x_j) - \frac{2}{nm}\sum_{i,j=1}^{n,m} k(x_i,y_j) + \frac{1}{m^2}\sum_{i,j=1}^{m} k(y_i,y_j).$$

A major drawback of the MMD is its computational cost : the complexity of computing $\mathrm{MMD}_k(\hat{\mathbb{P}},\hat{\mathbb{Q}})$ scales at least as $\Theta\left((n+m)^2 d\right)$, where d is the dimension of the data and it is assumed that the computational cost of evaluating the kernel k is $\Theta(d)$, which reduces to $\Theta\left(n^2 d\right)$, assuming n and m are of the same order. Methods to reduce this computational burden include linear MMD [GBR+12a], block MMD [ZGB13], and RFF MMD [ZM15]. In [GBR+12a], a simple linear time approximation by computing the MMD on a randomly chosen subset of $\sqrt{n}$ data points is used to deal with the *two-sample test* problem, a statistical test to assess whether two random samples share the same probability distribution. The reduced computational cost of $\Theta\left(nd\right)$ comes at the price of poor approximation properties. An improvement to this method was proposed in [ZGB13], where instead of considering a single subset of size $\sqrt{n}$, the average of the MMD on several blocks, each of size $s \leq n$, is computed. The number of blocks is usually equal to n/s and a commonly used heuristic for the block size is $\sqrt{n}$, leading to a complexity of $\Theta(n^{1.5}d)$. From a more general perspective, [ZM15] proposed the use of random Fourier Features (RFF) [RR07a] to approximate the (translation invariant) kernel function and then efficiently compute the MMD. The computational complexity is in this case reduced to $\Theta(nsd)$, where $s \ll n$ is the number of random features. It is worth noting that one major obstacles towards the use of kernel approximation techniques, such as Nyström [WS00], is that the embedding defined by the approximate kernel representations can no longer be injective, whereas it is crucial for the use of kernel mean embeddings to be theoretically supported.

2.2 Discrete RKHS

We here recall the core of discrete RKHSs. For a complete description, see [JT15].

Definition 1. *(Positive (semi-)definite kernel on a discrete domain)*

Let V be a countable and possibly infinite set, and $\mathcal{F}(V)$ the set of all finite subsets of V. A function $k : V \times V \to \mathbb{R}$ is positive semi-definite (psd)*, if it is symmetric and for all F in $\mathcal{F}(V)$ and all sets of real coefficients $\{c_x \in \mathbb{R}\}_{x\in F}$:*

$$\sum_{(x,y)\in F\times F}\sum c_x c_y k(x,y) \geq 0. \tag{1}$$

If for all $F \in \mathcal{F}(V)$ equality holds in (1) *implies that all elements of $\{c_x\}_{x\in F}$ are zero then k is said to be* positive definite (pd)*. It is said to be λ-*pd *if the right-hand side of* (1) *is replaced by $\lambda\|c\|^2$ for some $\lambda > 0$.*

Definition 2. *(RKHS on a discrete domain)*

Let V be a countable and possibly infinite set. A Hilbert space $\mathcal{H}$ of functions from V to $\mathbb{R}$ is a reproducing kernel Hilbert space (RKHS) *if there is a positive semi-definite kernel $k : V \times V \to \mathbb{R}$ such that :*

(i) the function $k_x := k(\cdot, x) : V \to \mathbb{R}$ belongs to $\mathcal{H}$ for all $x \in V$,

(ii) for every $\varphi \in \mathcal{H}$ and $x \in V$,

$$\langle k_x, \varphi \rangle = \varphi(x). \tag{2}$$

On account of (ii), the kernel k is called the reproducing kernel of $\mathcal{H}$. There is a natural bijection between positive semi-definite kernels on a given set V, and RKHS of functions on that set [Aro50].

Definitions 1 and 2 are not new and it is well-known that the theory of reproducing kernels is valid for discrete and continuous domains. Kernel methods have for instance already been applied with success to deal with structured data such as sequences, trees or graphs [Gar08]. What is new, however, is the following notion of discrete RKHSs, which characterizes RKHSs on discrete domains having the *discrete mass property.*

Definition 3. *(Discrete mass property)*

The RKHS $\mathcal{H}$ of functions defined on a countable infinite discrete set V is said to have the discrete mass property (and $\mathcal{H}$ is called a discrete RKHS), if $\delta_x \in \mathcal{H}$, for all $x \in V$, where δ_x is the Dirac mass at x, i.e., $\delta_x(y) = \begin{cases} 1 & \text{if } x = y, \\ 0 & \text{otherwise.} \end{cases}$

The next theorem gives a necessary and sufficient condition characterizing which point-masses from V are in $\mathcal{H}$.

Theorem 1. *(Necessary and sufficient condition for $\delta_x \in \mathcal{H}$)*

Let k be a psd *kernel defined on a countable and possibly infinite set V, and let $\mathcal{H}$ be the corresponding RKHS. Let $x \in V$ be given, and $\mathcal{F}(V)$ be the set of all finite subsets of V. Then $\delta_x \in \mathcal{H}$ if and only if*

$$\sup_{\{F \in \mathcal{F}(V) : x \in F\}} (K_F^{-1})_{x,x} < \infty, \tag{3}$$

where $K_F := (k(x, y))_{(x,y) \in F^2}$ is the Gram matrix of k on F.

In this case, we have :

$$\|\delta_x\|_{\mathcal{H}}^2 = \sup_{\{F \in \mathcal{F}(V) : x \in F\}} (K_F^{-1})_{x,x}.$$

This is an important result, as it states that the Dirac mass δ_x is in an RKHS when condition (3) is satisfied, which is required to fully leverage the advantages provided by reproducing kernels in the discrete case. Given this, it is possible to obtain an exact characterization of the orthogonal projection of the Dirac masses onto the span of the kernel functions.

Lemma 2. *(Orthogonal projection of Dirac masses onto $\mathcal{H}_F$)*

Let k, $\mathcal{H}$ be as above and F in $\mathcal{F}(V)$. Set $\mathcal{H}_F :=$ Span $(\{k_x\}_{x \in F})$. Let $P_F : \mathcal{H} \to \mathcal{H}_F$ be the orthogonal projection onto $\mathcal{H}_F$. For $x \in F$ such that $\delta_x \in \mathcal{H}$, we have

$$P_F(\delta_x) = \sum_{y \in F} (K_F^{-1})_{x,y} k_y. \tag{4}$$

In the following, we only consider the case where the kernel k is λ-pd (Def. 1), which means that the spectrum of K_F is in $[\lambda, +\infty)$, for any $F \in \mathcal{F}(V)$. In this case, $(K_F^{-1})_{x,x} \leq \lambda^{-1}$ for all F and x. Thus, condition (3) is always satisfied and all the Dirac masses δ_x belong to the RKHS associated to k. It is useful to point out that these assumptions are not restrictive. For example, for any positive semi-definite kernel $\tilde{k}$, pick $\lambda > 0$ a regularization parameter, then the kernel k defined by $k(x, y) := \tilde{k}(x, y) + \lambda \delta_x(y), \forall x, y \in V$, satisfies these assumptions.

3 Comparing Probability Distributions in Discrete RKHS

Dirac masses being elements of an RKHS provides a natural way to deal with such distributions. Indeed, given a probability distribution $\mathbb{P}$ that is only accessible through discrete and finite samples $X = \{x_i\}_{i=1}^n$, the corresponding empirical distribution can be written as $\hat{\mathbb{P}} = \sum_{x \in X} p_x \delta_x$, where δ_x is the Dirac mass at point x and p_x is the probability mass associated to the sample x. So, when the δ_x's are in an RKHS $\mathcal{H}$, $\hat{\mathbb{P}}$ also belongs to $\mathcal{H}$ and can be manipulated with the artillery of RKHSs to tackle probabilistic modeling problems (see Figure 1).

We now turn our attention to defining a kernel-based distance between empirical distributions. In the case where these are in the discrete RKHS $\mathcal{H}$, we can build distances using the inner product of $\mathcal{H}$. More formally, let V be a countable and possibly infinite set, $X = \{x_i\}_{i=1}^n$, $Y = \{y_i\}_{i=1}^m$ and $\hat{\mathbb{P}} = \sum_{x \in X} p_x \delta_x$ and $\hat{\mathbb{Q}} = \sum_{y \in Y} q_y \delta_y$ two discrete probability distributions. When the Dirac masses are elements of the RKHS $\mathcal{H}$,

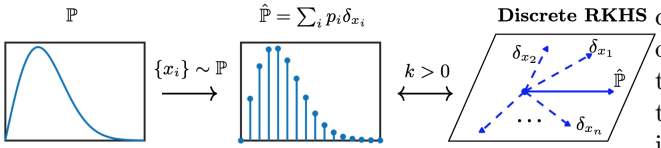


FIGURE 1 – Characterizing probability distributions in discrete RKHS. Given a probability distribution $\mathbb{P}$ that is only accessible through discrete and finite samples $X = \{x_i\}_{i=1}^n$, the corresponding empirical distribution can be written as $\hat{\mathbb{P}} = \sum_i p_i \delta_{x_i}$, where δ_{x_i} is the Dirac function at point x_i and p_i is the probability mass associated to the sample x_i. In contrast to "standard" RKHS, a *discrete RKHS* $\mathcal{H}$ contains the Dirac masses δ_{x_i}. $\hat{\mathbb{P}}$ is a linear combination of δ_{x_i}, and thus belongs to $\mathcal{H}$. A kernel $k > 0$ (i.e., positive definite) is needed to ensure that $\delta_{x_i} \in \mathcal{H}$.

so are the discrete probability distributions. We can then simply define the distance between $\hat{\mathbb{P}}$ and $\hat{\mathbb{Q}}$ as $\|\hat{\mathbb{P}} - \hat{\mathbb{Q}}\|_{\mathcal{H}}$. If it is sound, it cannot be computed explicitly as it requires the evaluation of the dot products $\langle \delta_x, \delta_y \rangle_{\mathcal{H}}$ which can not be evaluated through kernel computations. Indeed, the inner product between two functions in a RKHS is not known in general, however we can compute the inner product with the kernel function To remedy this problem, we introduce a linear operator O_k that maps the Dirac masses into the span of the kernel functions $\text{Span}\left(\{k_z\}_{z\in V}\right)$ and define a distance between $\hat{\mathbb{P}}$ and $\hat{\mathbb{Q}}$ parameterized by O_k. More formally, if for V, we let $\mathcal{D}_V := \text{Span}\left(\{\delta_z\}_{z\in V}\right)$ be the span of the Dirac masses associated with the elements of V, we propose the following definition.

Definition 4. *(Discrete RKHS distance between distributions)*
Let $\mathcal{H}$ be a discrete RKHS of functions defined on a countable infinite discrete set V and k its reproducing kernel. Let $O_k : \mathcal{D}_V \to \text{Span}\left(\{k_z\}_{z\in V}\right)$ be a linear operator with $\text{Null}(O_k) = \{0\}$. We define the Discrete RKHS distance D_{O_k} between two discrete probability distributions $\hat{\mathbb{P}}$ and $\hat{\mathbb{Q}}$ by

$$D_{O_k}(\hat{\mathbb{P}}, \hat{\mathbb{Q}}) := \|\hat{\mathbb{P}} - \hat{\mathbb{Q}}\|_{O_k} = \|O_k(\hat{\mathbb{P}} - \hat{\mathbb{Q}})\|_{\mathcal{H}}. \quad (5)$$

D_{O_k} is the distance induced by $\langle \cdot, \cdot \rangle_{O_k} := \langle O_k(\cdot), O_k(\cdot) \rangle_{\mathcal{H}}$ defined on $\mathcal{D}_V \times \mathcal{D}_V$, which *is* a dot product : it is a symmetric and a semi-definite positive bilinear form thanks to the linearity of O_k ; it is also definite positive ($\langle u, u \rangle_{O_k} \geq 0, \forall u$ and $\langle u, u \rangle_{O_k} = 0 \Rightarrow u = 0$) because the null space of O_k is reduced to zero.

The set of linear operators with zero null-space thus defines a family of distances. A question arises : how can we choose O_k ? In the following, we exhibit a family of operators O_k that are based on orthogonal projections, which ensures that this distances can be computed using the kernel. Interestingly, the MMD distance is one particular instance in this family. We leave the study of other possible choices of O_k for future work.

O_k maps $\mathcal{D}_V$, which is a subset of the discrete RKHS $\mathcal{H}$, to $\text{Span}\left(\{k_z\}_{z\in V}\right)$, just as orthogonal projections defined in Lemma 2. Using these projections is then a natural way to build O_k. Here, we consider a weighted sum of the projections on each feature k_z, $P_{\{z\}}(u)$, for all z in V, i.e.,

$$O_k : \mathcal{D}_V \to \text{Span}\left(\{k_z\}_{z\in V}\right) \quad (6)$$
$$u \mapsto O_k(u) := \sum_{z\in V} o_z P_{\{z\}}(u), \text{with } o_z \neq 0 \ \forall z.$$

Before going further, we show that O_k in (6) is a well-defined linear operator on $\mathcal{D}_V$, i.e., that the infinite sums $\sum_{z\in V} P_{\{z\}}(u)$ are convergent. In fact, we compute the value of $O_k(u)$:

Lemma 3. *Let V, k and $\mathcal{H}$ be as above, we have*
(i) $(y, z) \in V^2$: $P_{\{y\}}(\delta_y) = \frac{1}{k(y,y)} k_y$ and if $z \neq y$, $P_{\{z\}}(\delta_y) = 0$.
(ii) $u := \sum_{i=1}^n \alpha_i \delta_{y_i} \in \mathcal{D}_V \Rightarrow O_k(u) = \sum_{i=1}^n \frac{\alpha_i o_{y_i}}{k(y_i,y_i)} k_{y_i} \in \text{Span}\left(\{k_z\}_{z\in V}\right)$.

proof. *For (i) : $P_{\{z\}}(\delta_y) = \frac{\langle \delta_y, k_z \rangle_{\mathcal{H}}}{\|k_z\|^2_{\mathcal{H}}} k_z = \frac{\delta_y(z)}{k(z,z)} k_z$. For (ii) : from (i), the infinite sum $\sum_{z\in V} o_z P_{\{z\}}(\delta_y)$ reduces to one term and $O_k(\delta_y) = o_y P_{\{y\}}(\delta_y) = \frac{o_y}{k(y,y)} k_y$ for all y. Similarly for $u = \sum_{i=1}^n \alpha_i \delta_{y_i}$, $P_z(u) = \frac{\alpha_z o_z}{k(z,z)} \delta_z$ if $z \in \{y_i\}_{i=1}^n$, and $P_z(u) = 0$ otherwise. The infinite sum in $O_k(u)$ reduces to n terms, it is finite thus convergent, and the result holds.*

To make sure that D_{O_k} is a metric in $\mathcal{D}_V$, it suffices to check that $\text{Null}(O_k) = \{0\}$. Given that

$$\|O_k(u)\|^2_{\mathcal{H}} = \left\| \sum_i \frac{\alpha_i o_{y_i}}{k(y_i, y_i)} k_{y_i} \right\|^2_{\mathcal{H}} = \sum_{i,j} \frac{\alpha_i o_{y_i}}{k(y_i, y_i)} \frac{\alpha_j o_{y_j}}{k(y_j, y_j)} k(y_i, y_j) \quad (7)$$

and that we have assumed that k is positive definite,

$$O_k(u) = 0 \Rightarrow \frac{\alpha_i o_{y_i}}{k(y_i, y_i)} = 0 \ \forall i,$$

with the o_y being nonzero, it yields $\alpha_i = 0 \ \forall i$ i.e. $u = 0$. Thus $\text{Null}(O_k) = \{0\}$.

This says any operator O_k as in (6) defines a distance D_{O_k} on $\mathcal{D}_V$. A key observation is that maximum mean discrepancy (MMD) is an instance of D_{O_k}.

Theorem 4. *Let V, k, $\mathcal{H}$ be as above and $O_k : u \in \mathcal{D}_v \mapsto \sum_{z\in V} k(z,z)P_{\{z\}}(u)$. Let $\hat{\mathbb{P}} := \sum_{x\in X} p_x\delta_x$ and $\hat{\mathbb{Q}} := \sum_{y\in Y} q_y\delta_y$ be two discrete probability distributions. We have*

$$D_{O_k}(\hat{\mathbb{P}},\hat{\mathbb{Q}}) = \mathrm{MMD}_k(\hat{\mathbb{P}},\hat{\mathbb{Q}}).$$

proof. *Since for all $k(z,z) \neq 0$, O_k defined above fits (6) and D_{O_k} is a distance. It is also clear that $O_k(\hat{\mathbb{P}}) = \sum_{x\in X} p_x k_x$ so O_k maps a discrete distribution $\hat{\mathbb{P}}$ (in $\mathcal{D}_V \subset \mathcal{H}$) to its mean embedding $\hat{\mu}_{\mathbb{P}}$ (in $\mathcal{H}$) so that $D_{O_k}(\hat{\mathbb{P}},\hat{\mathbb{Q}}) = \|O_k(\hat{\mathbb{P}}-\hat{\mathbb{Q}})\|_{\mathcal{H}} = \|\hat{\mu}_{\mathbb{P}} - \hat{\mu}_{\mathbb{Q}}\|_{\mathcal{H}} = \mathrm{MMD}_k(\hat{\mathbb{P}},\hat{\mathbb{Q}})$.*

Theorem 4 shows how MMD is a particular instance of the proposed discrete RKHS distances defined in Def 4. From a different perspective, it is interesting to note that our formulation provides a new understanding to the RKHS characterization of probability distributions and gives theoretical justification for the use of MMD with a large class of kernels, as soon as they satisfy the positive definiteness assumption and condition (3). This also complements the literature of RKHS embedding of measures [SGS18, SFL11]. In the following we propose a fast approximation of MMD based on the Nyström method and discrete RKHS.

4 A Nyström-based MMD Approximation

The Nyström method is among the most used techniques for approximating the kernel Gram matrix of a large data sample [WS00], building an approximate Gram matrix based on a small subset of the training points, called landmarks.

Consider a training set $F := \{x_i\}_{i=1}^n$ of n training samples.To approximate the kernel matrix $K_F \in \mathbb{R}^{n\times n}$, the Nyström method randomly samples $s \ll n$ examples $S := \{\hat{x}_i\}_{i=1}^s \subset F$ and forms the $n \times s$ matrix $K_{F,S} := (k(x,x'))_{(x,x')\in F\times S}$, and the (small) $s \times s$ kernel matrix $K_S := (k(x,x'))_{(x,x')\in S^2}$. The Nyström approximation is obtained as follows

$$\hat{K}_F := K_{F,S}K_S^+K_{F,S}^\top \approx K,$$

where K_S^+ denotes the pseudo-inverse of K_S. The feature representation associated to the Nyström method is given by :

$$\hat{\phi}(x) := \left([k(\hat{x}_1,x),\ldots,k(\hat{x}_s,x)]\,(K_S^+)^{\frac{1}{2}}\right)^\top,$$

and the approximated kernel function is $\hat{k}_{nys}(x,x') := \langle\hat{\phi}(x),\hat{\phi}(x')\rangle$. It is straightforward to verify that, for all $x_i,x_j \in F$, $\langle\hat{\phi}(x_i),\hat{\phi}(x_j)\rangle = (\hat{K}_F)_{ij}$.

To take advantage of the framework of discrete RKHS, we consider the kernel $\hat{k}$ defined on a countable and infinite set V and obtained by regularizing $\hat{k}_{nys}$, i.e., $\forall x,y \in V$, $\hat{k}(x,y) = \hat{k}_{nys}(x,y) + \lambda\delta_x(y)$, where $\lambda > 0$ is a regularization parameter (see end of Section 2). The following lemma shows that the MMD distance using the kernel $\hat{k}$ instead of k also defines a metric between discrete probability distributions.

Lemma 5. *Let V and k be as above. If $\hat{k}$ is a regularized version of a Nyström approximation of the kernel function k : $\hat{k}(x,y) = \hat{k}_{nys}(x,y) + \lambda\delta_x(y)$, for all $x,y \in V$ and with $\lambda > 0$, then $\mathrm{MMD}_{\hat{k}}(\cdot,\cdot)$ is a metric on $\mathcal{D}_V$.*

proof. *This is a direct application of Theorem 4.*

This result justifies the use of Nyström approximation with MMD, which is an interesting result in its own right. We now show that this translates to a significant computational saving compared to the $\Theta(n^2d)$ cost of exact MMD using the kernel k.

Lemma 6. *Let V and k be as above. Let F be a finite subset of V of size n, and $\hat{k}$ be a regularized Nyström approximation of k using s landmarks. For any two discrete probability distributions $\hat{\mathbb{P}} := \sum_{x\in F} p_x\delta_x$ and $\hat{\mathbb{Q}} := \sum_{y\in F} q_y\delta_y$, the computational complexity of $\mathrm{MMD}_{\hat{k}}(\hat{\mathbb{P}},\hat{\mathbb{Q}})$ is $\Theta(nsd)$.*

proof. $\mathrm{MMD}^2_{\hat{k}}(\hat{\mathbb{P}},\hat{\mathbb{Q}}) = \sum_{x,y\in F}(p_x-q_x)\hat{k}(x,y)(p_y-q_y) = \mathrm{v}K_{F,S}K_S^+K_{F,S}^\top\mathrm{v}^\top + \lambda\mathrm{r}\mathrm{v}^\top$, *with* v *the row vector* $\mathrm{v} = (p_x-q_x)_{x\in F}$. *This computation requires $\Theta(ns)$ operations once the matrices $K_{F,S}$ and K_S^+ are known. Building these matrices requires $\Theta(nsd)$ operations.*

We now address the question : how far is $\mathrm{MMD}_{\hat{k}}$ from MMD_k ? To answer this question, we make use of recent advances on the performance quality of sampling methods on kernel matrices [GM16]. In the case of uniform sampling, assumptions about the coherence properties of the kernel matrix are required. The coherence of the top r-dimensional eigenspace of a $n \times n$ kernel matrix K, denoted by μ_r, is defined as : $\mu_r := \frac{n}{r}\max_{i\in\{1,\ldots,n\}}\|U_i\|^2$, where U is the $n \times r$ matrix containing the top r-eigenvectors of K.

Lemma 7. *Let V, k and μ_r be as above. Let F be a finite subset of V of size n, and $\hat{k}$ be a regularized Nyström approximation of k using s landmarks sampled uniformly at random from F and a regularization*

parameter $\lambda > 0$. Fix a failure probability $\delta \in (0,1)$ and an accuracy factor $\varepsilon \in (0,1)$. For any two discrete probability distributions $\hat{\mathbb{P}} := \sum_{x \in F} p_x \delta_x$ and $\hat{\mathbb{Q}} := \sum_{y \in F} q_y \delta_y$, if $s \geq 2\mu\varepsilon^{-2} r \ln(r/\delta)$, then it holds, with probability at least $1 - 3\delta$, that

$$\left|\mathrm{MMD}_k^2(\hat{\mathbb{P}}, \hat{\mathbb{Q}}) - \mathrm{MMD}_{\hat{k}}^2(\hat{\mathbb{P}}, \hat{\mathbb{Q}})\right| \leq (2 + \frac{2}{\delta^2(1-\varepsilon)})\|K - K_r\|_* + 2\lambda,$$

where K_r is the best rank-r approximation of the kernel matrix K, and $\|\cdot\|_*$ denotes the nuclear norm.

proof. *Let* v *be the row vector $(p_x - q_x)_{x \in F}$.*

$$\begin{aligned}
&\left|\mathrm{MMD}_k^2(\hat{\mathbb{P}}, \hat{\mathbb{Q}}) - \mathrm{MMD}_{\hat{k}}^2(\hat{\mathbb{P}}, \hat{\mathbb{Q}})\right| \\
&= \left|\mathrm{Tr}(K\mathrm{r}^\top \mathrm{v}) - \mathrm{Tr}(\hat{K}\mathrm{v}^\top \mathrm{v})\right| \\
&\leq \left|\mathrm{Tr}\big((K - K_{F,S}K_S^+K_{F,S}^\top)\mathrm{v}^\top \mathrm{v}\big)\right| + \lambda\|\mathrm{v}\|^2 \\
&\overset{(a)}{\leq} \|K - K_{F,S}K_S^+K_{F,S}^\top\|_* \|\mathrm{v}^\top \mathrm{v}\|_\infty + \lambda\|\mathrm{v}\|^2 \\
&\overset{(b)}{\leq} (1 + \frac{1}{\delta^2(1-\varepsilon)})\|K - K_r\|_*\|\mathrm{v}\|^2 + \lambda\|\mathrm{v}\|^2,
\end{aligned}$$

where inequality (a) follows from Hölder's inequality [Bha97], and inequality (b) follows from [GM16, Lemma 8]. Since $\sum_{x \in F} p_x = \sum_{y \in F} q_y = 1$ and $0 \leq p_x, q_x \leq 1$, we have $\|\mathrm{v}\|^2 \leq 2$, which completes the proof.

Note that the quality of the bound can be improved when better sampling strategies, such as leverage score sampling, are used [GM16].

5 Experiments

We now turn our attention to empirically evaluate the Nyström based MMD approximation. We conduct experiments on the three sample problem in a simulated setting and on real data. For reproducibility, our code will be made publicly available.

The three-sample problem consists in the following : given three samples $X = \{x_i\}_{i=1}^{n_x}$, $Y = \{y_i\}_{i=1}^{n_y}$ and $Z = \{z_i\}_{i=1}^{n_z}$ such that X and Y are generated from two different distributions ($\mathbb{P}_X$ and $\mathbb{P}_Y$), identify whether Z is generated from the same distribution as X or Y [Gut89, RM13] —where it is assumed, of course, that Z is generated from one of these two distributions. This problem can be addressed using a suitable distance that measures the similarity between probability distributions. In other words, given a distance D, we decide that Z is generated from $\mathbb{P}_X$ if $D(\mathbb{P}_Z, \mathbb{P}_X) < D(\mathbb{P}_Z, \mathbb{P}_Y)$. Otherwise, we decide that $Z \sim \mathbb{P}_Y$. Compared to the two-sample problem [GBR+12a], which aims at answering the question whether two samples are identically distributed, no threshold is needed to make the decision.

Dataset	Size	Type
$\mathcal{N}(0,1)$ and $\mathcal{N}(0,2)$	30 000	synthetic
Fast food [1]	10 000	real-world
MUTAG [DLdCD+91]	188	real-world

TABLE 1 – Datasets.

5.1 Experimental Setup

We consider two different structures of data : vectors and graphs, that we summarize in Table 1. For vector data, we used the classical RBF kernel with parameter $\sigma = \left(\frac{1}{n}\sum_{x_i,x_j \in X} \|x_i - x_j\|_2^2\right)^{-1}$ as proposed in [BGSW06], while for graph data we used the ShortestPath kernel [BK05]. We only consider the case where X, Y and Z have the same number of data, but our framework is still valid if they have different sizes.

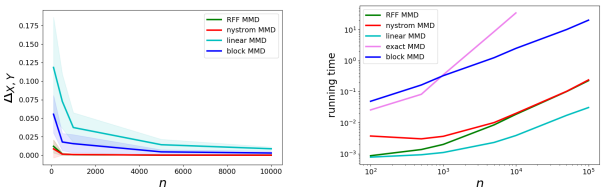
We compare our Nyström based MMD approximation (Nyström MMD) with three other methods : the Random Fourier Features based MMD (RFF MMD) [AKM+17], the block MMD [ZGB13] and the linear approximation (linear MMD) [GBR+12a]. When possible, we also compute the exact value of the MMD (exact MMD). Remember that the complexity of the computation is in $\Theta(n^2 d)$ for exact MMD, $\Theta(nd)$ for linear MMD, and $\Theta(nsd)$ with s the number of data sampled for Nyström MMD, s the number of Fourier vectors sampled for RFF MMD and s the size of the blocks for block MMD. To compare methods of equivalent complexity, we fix the value s to $\log(n)$ for all three methods.

We are interested in the error in solving the three-sample problem computed as the fraction of test sampled misplaced : $error = \frac{|\{z \in Z : z \sim \mathbb{P}_Y\}|}{|Z|}$. We also compute the running time. To further evaluate the quality of the MMD approximations, we compute the distance of the approximated MMD to the true value :

$$\Delta_{X,Y} := |\mathrm{MMD}_k(X,Y) - \mathrm{MMD}_{\hat{k}}(X,Y)|,$$

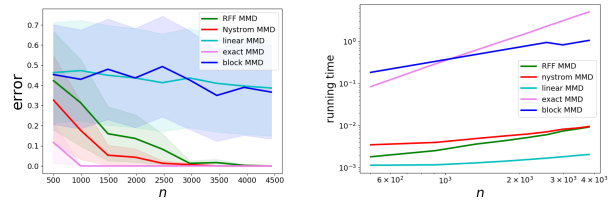
where $\mathrm{MMD}_{\hat{k}}$ is computed using one of the approximation methods : block MMD, RFF MMD, linear MMD

1. https ://www.kaggle.com/datasets/datafiniti/fast-food-restaurants



(a) distance to exact MMD (b) running time

FIGURE 2 – Three sample problem on $\mathcal{N}(0,1)$ and $\mathcal{N}(0,2)$ data.



(a) error (b) running time

FIGURE 3 – Three sample problem on the the Fast food dataset.

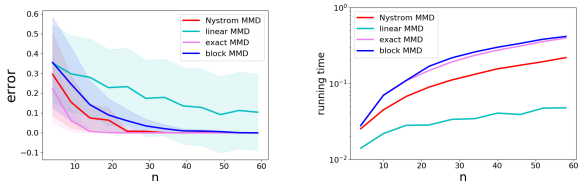
or Nyström MMD.

All experiments are run several times and the quantities in the following results are averaged all the runs.

5.2 Experimental Results

We first consider a synthetic dataset, consisting in 5000 samples for X, Y and Z and with $\mathbb{P}_X = \mathbb{P}_Z = \mathcal{N}(0,1)$ and $\mathbb{P}_Y = \mathcal{N}(0,2)$. The experiment is run 10 times. The three-sample problem in this case is pretty simple and all methods achieve a null error. The interest here lies in confirming that the actual running time scales with the complexity, and that the MMD approximation are good approximation. As shown in Figure 2(b) exact MMD has an exploding running time while the RFF and Nyström MMD have the same running time, and are faster than block-MMD, but slower than linear MMD. This is in line with the expected behaviour of these methods. We plot in Figure 2(a) the distance of the approximated MMD to the true value, $\Delta_{X,Y}$, as a function of number of data n (and its variance in lighter colors). This plot shows that the Nyström and the RFF method have a similar accuracy, and are more efficient at approximating the true value of MMD than block MMD and linear MMD.

We now consider the FastFood dataset which contains the localisation of some restaurants in America. The whole dataset contains initially 600 classes.



(a) error (b) running time

FIGURE 4 – Three-sample problem on the MUTAG dataset.

For this experiemnts we group them into 2 classes of roughly 5000 points each. The datasets X,Y and Z then contains about 2500 points in each of the 100 runs. We plot the results of these experiments in Figure 3. The error in the three-sample problem Figure 3(a) and the running time in Figure 3(b) confirm those found in the previous experiments. We can see that the Nyström MMD obtains a lower error rate compared to the other MMD approximation methods.

Let now consider our last dataset, MUTAG [DLdCD+91], which is a graph dataset for binary classification. We take X and Z from the set of graphs that are labelled $+1$, and Y from the set of graphs that are labelled -1. We have about $n = 60$ graphs in each set X, Y and Z. A useful kernel in this case is the ShortestPath kernel [BGSW06]. We make the same experiments as above, and plot the results in Figure 4 where the error and running time are shown as a function of n and with $s = \log(n)$ and are averaged over 1000 runs. The RFF MMD method is not suitable for graph kernels. Since the number of data is small, one can compute exact MMD. In terms of running time, Nyström MMD is faster than exact and block MMD. Linear MMD is faster than Nyström MMD but suffers from a much higher error rate. Nyström MMD achieves a good trade-off between effectiveness and efficiency.

6 Conclusion

We uncovered a new way to make the connection between probability distributions and kernel methods, using tools of discrete RKHSs. It allowed us to come up with a novel framework for representing and comparing probability distributions, and to show that the MMD distance can be retrieved from this framework. Pushing the envelope further, we proposes a new fast approximation to the MMD distance, based on the Nyström method. Our MMD approximation is theoretically jus-

tified for a large class of kernels, including graph kernels.

We finally have empirically evaluated our Nyström MMD on both simulated and real world datasets. We have shown that our MMD approximation achieves a good trade-off between error and running time in solving the three-sample problem.

Future work would focus on further studying the versatility of our framework and see how other distances between distributions, such as Wasserstein distances, fall into its realm.

Références

[ABR64] Mark A Aizerman, È. M. Braverman, and L. I. Rozonoèr. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and remote control*, 25 :821–837, 1964.

[AKM+17] Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, and Amir Zandieh. Random fourier features for kernel ridge regression : Approximation bounds and statistical guarantees. In *International Conference on Machine Learning*, pages 253–262. PMLR, 2017.

[Aro50] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3) :337–404, 1950.

[BGSW06] Ulf Brefeld, Thomas Gärtner, Tobias Scheffer, and Stefan Wrobel. Efficient co-regularised least squares regression. In *Proceedings of the 23rd international conference on Machine learning*, pages 137–144, 2006.

[BGV92] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[Bha97] Rajendra Bhatia. *Matrix analysis*. Springer, 1997.

[BK05] Karsten M. Borgwardt and Hans-Peter Kriegel. Shortest-path kernels on graphs. In *Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM 2005)*, pages 74–81, Washington, DC, USA, 2005. IEEE Computer Society.

[CFD+13] Rita Chattopadhyay, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. Joint transfer and batch-mode active learning. In *International conference on machine learning*, pages 253–261. PMLR, 2013.

[CFTR16] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9) :1853–1865, 2016.

[CG95] S. L. Campbell and C. W. Gear. The index of general nonlinear DAES. *Numer. Math.*, 72(2) :173–196, 1995.

[CMT10] Corinna Cortes, Mehryar Mohri, and Ameet Talwalkar. On the impact of kernel approximation on learning accuracy. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 113–120. JMLR Workshop and Conference Proceedings, 2010.

[CRSG15] Kacper P Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. *Advances in Neural Information Processing Systems*, 28 :1981–1989, 2015.

[CWF+13] Rita Chattopadhyay, Zheng Wang, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. Batch mode active sampling based on marginal probability distribution matching. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 7(3) :1–25, 2013.

[DLdCD+91] Asim Kumar Debnath, Rosa L Lopez de Compadre, Gargi Debnath, Alan J Shusterman, and Corwin Hansch. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of medicinal chemistry*, 34(2) :786–797, 1991.

[DMC05] Petros Drineas, Michael W Mahoney, and Nello Cristianini. On the nyström

method for approximating a gram matrix for improved kernel-based learning. *journal of machine learning research*, 6(12), 2005.

[FBJ04] Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5(Jan) :73–99, 2004.

[FS01] Shai Fine and Katya Scheinberg. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2(Dec) :243–264, 2001.

[Gar08] Thomas Gartner. *Kernels for structured data*, volume 72. World Scientific, 2008.

[GBR+12a] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1) :723–773, 2012.

[GBR+12b] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1) :723–773, 2012.

[GHLM13] Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A pac-bayesian approach for domain adaptation with specialization to linear classifiers. In *International conference on machine learning*, pages 738–746. PMLR, 2013.

[GM16] Alex Gittens and Michael W Mahoney. Revisiting the nyström method for improved large-scale machine learning. *The Journal of Machine Learning Research*, 17(1) :3977–4041, 2016.

[GSH+09] Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4) :5, 2009.

[Gut89] Michael Gutman. Asymptotically optimal classification for multiple tests with empirically observed statistics. *IEEE Transactions on Information Theory*, 35(2) :401–408, 1989.

[HC] Aric Hagberg and Drew Conway. Networkx : Network analysis with python.

[HSS08] Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. *The annals of statistics*, pages 1171–1220, 2008.

[JT15] Palle Jorgensen and Feng Tian. Discrete reproducing kernel hilbert spaces : sampling and distribution of dirac-masses. *The Journal of Machine Learning Research*, 16(1) :3079–3114, 2015.

[KNS+19] Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo K Rohde. Generalized sliced wasserstein distances. *arXiv preprint arXiv :1902.00434*, 2019.

[LCC+17] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan : Towards deeper understanding of moment matching network. *arXiv preprint arXiv :1705.08584*, 2017.

[LPSS+14] David Lopez-Paz, Suvrit Sra, Alex Smola, Zoubin Ghahramani, and Bernhard Schölkopf. Randomized nonlinear component analysis. In *International conference on machine learning*, pages 1359–1367. PMLR, 2014.

[MFSS16] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions : A review and beyond. *arXiv preprint arXiv :1605.09522*, 2016.

[MFSS17] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions : A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2) :1–141, 2017.

[RM13] Daniil Ryabko and Jérémie Mary. A binary-classification-based metric between time-series distributions and its use in statistical and learning problems. *The Journal of Machine Learning Research*, 14(1) :2837–2856, 2013.

[RR07a] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

[RR07b] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

[SFL11] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(7), 2011.

[SGS18] Carl-Johann Simon-Gabriel and Bernhard Schölkopf. Kernel distribution embeddings : Universal kernels, characteristic kernels and kernel metrics on distributions. *The Journal of Machine Learning Research*, 19(1) :1708–1736, 2018.

[SS00] Alex J Smola and Bernhard Schölkopf. Sparse greedy matrix approximation for machine learning. 2000.

[SS02] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels : support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[SS15] Danica J Sutherland and Jeff Schneider. On the error of random fourier features. *arXiv preprint arXiv :1506.02785*, 2015.

[STC04] John Shawe-Taylor and Nello Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[WS00] Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. *Advances in neural information processing systems*, 13, 2000.

[YLM+12] Tianbao Yang, Yu-Feng Li, Mehrdad Mahdavi, Rong Jin, and Zhi-Hua Zhou. Nyström method vs random fourier features : A theoretical and empirical comparison. *Advances in neural information processing systems*, 25 :476–484, 2012.

[ZGB13] Wojciech Zaremba, Arthur Gretton, and Matthew Blaschko. B-tests : Low variance kernel two-sample tests. *arXiv preprint arXiv :1307.1954*, 2013.

[ZM15] Ji Zhao and Deyu Meng. Fastmmd : Ensemble of circular discrepancy for efficient two-sample test. *Neural computation*, 27(6) :1345–1372, 2015.