A GENERALIZED PROBABILITY KERNEL ON DISCRETE DISTRIBUTIONS AND ITS APPLICATION IN TWO-SAMPLE TEST

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a generalized probability kernel on discrete distributions with finite support. This probability kernel, defined as kernel between distributions instead of samples, generalizes a variety of existing discrepancy statistics including maximum mean discrepancy(MMD) as well as kernelized Stein discrepancy(KSD), and extends to more general cases. For both existing and newly proposed statistics, we estimate them through empirical frequency and illustrate the strategy to search for the unbiased ones. We further analyze several unbiased plugin(empirical) estimators for the task of two-sample test. Our work connects the fields of discrete distribution-property estimation and kernel-based hypothesis test, which might shed light on more new possibilities.

1 INTRODUCTION

We focus on the two-sample problem, which is given two i.i.d samples $\{x_1, x_2, ...x_n\}, \{y_1, y_2, ..., y_n\}$, could we infer the descrepency between underline distributions they are drawn from. For such a problem, the option of hypothesis test i.e. two-sample test is most popular, and a variety of statistics in estimating the descrepency is proposed. In recent years, RKHS based method such as maximum mean discrepancy(MMD) has gained a lot of attention. Gretton et al. (2012) has shown that in a universal-RKHS F, $MMD(F, \boldsymbol{p}, \boldsymbol{q}) = 0$ if and only if $\boldsymbol{p} = \boldsymbol{q}$, thus could serve as a valid statistic for the two-sample hypothesis test. Gretton et al. (2012) further provides unbiased estimator of MMD with fast asymptotic convergance rate, illustrating its advantages. The work of MMD is then extended to the problem of goodness-of-fit, where only one sample $\{x_1, x_2, ...x_n\}$ drawn from distribution p is given and the other distribution q is given by a probability model. A modification of MMD through the usage of Stein operator results in kernelized Stein discrepancy(KSD) Qiang et al. (2016), and is then further extended to discrete setting Yang et al. (2018). All of the above methods have achieved great success which illustrates the advantages of RKHS-based methods over distribution free analysis.

On the other hand, estimating distribution properties with plugin(empirical) estimators on discrete setting is an active research area in recent years, where people focus on problem settings with large support size but not so large sample size. The Bernstein polynomial technique is introduced in analyzing the bias of the plugin estimators Yi & Alon (2020), which provides remarkable progress on bias-reduction methods of the plugin estimators. It is thus interesting to ask if the plugin estimators could motivate new results for the RKHS-based two-sample test.

Another interesting topic is about the probability kernel, defined as kernel function over probabilities, instead of over samples. As is easily seen, any discrepancy measure of distribution p and q could potentially be valid probability kernels, not so much work focuses on this. While Jebara et al. (2004) introduced the so called probability product kernels which generalize a variety of discrepancy measures, its properties remain further study.

Motivated by above observations, our work focuses on a specialized probability kernel function which is a direct generalization of sample-based RKHS methods such as MMD, KSD. We focus on using plugin-estimator as the default estimator of the kernel function we defined, and illustate that with the help of Bernstein polynomial techniques, we could analyze the bias of these plugin-

estimators. Our work thus connects the fields of discrete distribution-property estimation and kernel-based hypothesis test, which brings interesting possibilities.

2 OUR CONTRIBUTIONS

To summarize, our contribution is introducing a framework of the generalized probability kernel(GPK) which in general treats the function space of probability mass function as a subspace of universal-RKHS F. The resulting family of $GPK[F]$ is valuable especially when they are equiped with a plugin-estimator. We introduce a series of techniques in analyzing bias of these plugin-estimators and illustrate a procedure of searching for members of $GPK[F]$ family suitable for applications such as the two-sample test.

3 NOTATION

We use bold symbol $\boldsymbol{v}$ in representing a probability function over a discrete support, and $[k]$ in representing all the discrete values(labels) the samples can take $\{x_1, x_2, ..., x_k\}$, and symbol v_i in representing a probability measure of each label $p_i = p(x_i)$.

4 GENERALIZED PROBABILITY KERNEL

Probability kernel function, defined as kernel function between distributions instead of samples, is a natural extension of the idea of kernel function in sample space.

Definition 1. *Given distribution $\boldsymbol{p}$ and $\boldsymbol{q}$ belongs to a family of discrete distribution with the same finite suport in the function space $X \in \mathbb{R}^k$, where k is the support size, we define the probability kernel function as $K(p, q)$, which is a kernel function maps from $X \times X \in \mathbb{R}^k \times \mathbb{R}^k$ to real value $d \in \mathbb{R}$, where d indicates the discrepency between distributions.*

Many discrepency measures, such as MMD, can serve as probability kernel functions, but people usually don't use the term of probability kernel function when describing them. The reason is that for most of the time, we only consider a limited number of distributions, and do not need or have the resources to navigate through all the distributions within the family. For example, when looking into the two-sample problem, we usally assume two samples $\{x_1, x_2, ..., x_n\}$ and $\{y_1, y_2, ..., y_n\}$ are i.i.d drawn from two distributions p and q, and use the discrepency measure $MMD[F, p, q]$ to determine if p and q are indistinguishable in the RKHS F. We do not consider all other distributions in F that is irrelevant to our samples! So far the idea of kernel function between distributions is in practice not so much useful, however, here in this paper, we propose, when considering the plugin-estimator of many of the exsiting discrepency measures, it is beneficial to view them as probability kernel functions.

We firstly illustrate this idea through ananlyzing MMD from the viewpoint of probability kernel, and then introduce our definition of generalized probability kernel, which is a specialized probability kernel function with some interesting properties.

4.1 MMD AS PROBABILITY KERNEL

We directly start from the definition of squared MMD

By definition, MMD is an instance of integral probability metric for functions in certain RKHS $\mathcal{F}$.

$$
\begin{aligned}
\mathbb{MMD}^2[\mathcal{F}, p, q] &= \left[\sup_{\|f\|_{\mathcal{H}} \leq 1} \left(\mathbf{E}_x[f(x)] - \mathbf{E}_y[f(y)] \right) \right]^2 \\
&= \left[\sup_{\|f\|_{\mathcal{H}} \leq 1} \langle \mu_p - \mu_q, f \rangle_{\mathcal{H}} \right]^2 \\
&= \|\mu_p - \mu_q\|_{\mathcal{H}}^2
\end{aligned}
$$

The process of Gretton et al. (2012) takes advantage of reproducing property of RKHS, and realize a more widely used form via decomposing the product-sum form into sum-product.

$$\mathbb{MMD}^2[\mathcal{F}, p, q] = \mathbf{E}_{x,x'}\left[k\left(x, x'\right)\right] - 2\mathbf{E}_{x,y}[k(x, y)] + \mathbf{E}_{yy'}\left[k\left(y, y'\right)\right]$$

Where x and x' are independent random variables with distribution p, and y and y' are independent random variables with distribution q

The benefit of doing so is that we do not need to evaluate the embedding functions anymore. This is important since for some widely used universal kernels such as RBF kernel, their embedding functions are expensive or even intractable to be directly evaluated.

We next derive our probability kernel form of MMD. The intuition is that we can regroup sample summations according to each label, and result in a functional form of probability for each label under distribution p and q.

Recall

$$\begin{aligned}\mathbb{MMD}^2[\mathcal{F}, p, q] &= \|\mu_p - \mu_q\|_{\mathcal{H}}^2 \\ &= \|\mathbf{E}_{x\sim p}\phi(x) - \mathbf{E}_{x'\sim q}\phi(x')\|_{\mathcal{H}}^2\end{aligned}$$

Given discrete distribution with finite support, we can write the expectation exactly with distribution probability function $p(x), q(x)$

$$\begin{aligned}\mathbb{MMD}^2[\mathcal{F}, p, q] &= \|\mathbf{E}_{x\sim p}\phi(x) - \mathbf{E}_{x'\sim q}\phi(x')\|_{\mathcal{H}}^2 \\ &= \left\|\sum_{i\in[k]}\phi(x_i)p_i - \sum_{i\in[k]}\phi(x_i)q_i\right\|_{\mathcal{H}}^2 \\ &= \sum_{i\in[k]}\sum_{j\in[k]}(p_i - q_i)\phi(x_i)\phi(x_j)(p_j - q_j) \\ &= \sum_{i\in[k]}\sum_{j\in[k]}(p_i - q_i)K_{ij}(p_j - q_j)\end{aligned}$$

We are ready to define our generalized probability kernel function, which directly extend the above formula of MMD.

Definition 2 (Generalized probability kernel)**.** *Given the family S of discrete distribution on labels $\{y_1, y_2, .., y_n\}$, and a universal-RKHS F defined by kernel K, where $k(y_i, y_j)$ maps from $\mathbb{R}^n \times \mathbb{R}^n$ to $\mathbb{R}$, and a element-wise mapping functional on functions in the RKHS $\phi(\boldsymbol{p}, \boldsymbol{q})$, which maps from $\mathbb{R}^n \times \mathbb{R}^n$ to $\mathbb{R}^n$, the generalized probability kernel function on distribution $\boldsymbol{p}, \boldsymbol{q} \in S$ is $k_{prob}(\boldsymbol{p}, \boldsymbol{q}) = \phi\left(\boldsymbol{p}, \boldsymbol{q}\right) K \phi\left(\boldsymbol{p}, \boldsymbol{q}\right)^T = \sum_{i\in[k]}\sum_{j\in[k]}\phi(p_i, q_i)K_{ij}\phi(p_j, q_j)$*

More specifically, we are interested in GPK with a narrowed family of mapping functions.

Definition 3 (type I generalized probability kernel)**.** *A type I generalized probability kernel is a GPK that equiped with the mapping function ϕ that satisfied $\|\phi(\boldsymbol{p}, \boldsymbol{q})\| = 0$ if and only if $\boldsymbol{p} = \boldsymbol{q}$*

Obviously, MMD is the special case of this type I generalized probability kernel given $\phi(\boldsymbol{p}, \boldsymbol{q}) = \boldsymbol{p} - \boldsymbol{q}$

We use the term generalized probability kernel(GPK) as this family of kernel functions directly extends the case of MMD, which utilizes a RKHS in measuring the discrepancy between two distributions. In the following sections, we will show that other kernel based discrepancy measures such as KSD also belong to this family of probability kernel.

Below we firstly exam some of the very important properties of MMD through the GPK point of view, and then illustrate that these properties could be analyzed in the same way for more memebers in the GPK family.

4.2 MMD AS METRIC MEASURE

One of the reasons why MMD is suitable as metric for two-sample test is due to the theorem according to Gretton et al. (2012)

Theorem 1. *Let F be a unit ball in a universal RKHS H, defined on the compact metric space X, with associated continuous kernel $k(\cdot,\cdot)$. Then $MMD[F,\boldsymbol{p},\boldsymbol{q}]=0$ if and only if $\boldsymbol{p}=\boldsymbol{q}$.*

Gretton et al. (2012) have proved this theorem using the universal properties of RKHS, below we provide a more brief proof.

Proof. recall

$$\begin{aligned}\mathbb{MMD}^2[\mathcal{F},\boldsymbol{p},\boldsymbol{q}] &= \sum_{i\in[k]}\sum_{j\in[k]}(p_i-q_i)K_{ij}(p_j-q_j)\\ &= \sum_{i\in[k]}\sum_{j\in[k]}v_iK_{ij}v_j\\ &= \boldsymbol{v}K\boldsymbol{v}^T\end{aligned}$$

where $\boldsymbol{v}=\boldsymbol{p}-\boldsymbol{q}$ is also a function in RKHS. Since K in an universal kernel, it is a positive definite matrix in discrete setting. Thus by definition of positive definite matrix, $\boldsymbol{v}K\boldsymbol{v}^T\geq 0$, where eqaulity holds if and only if $\boldsymbol{v}=\mathbf{0}$, and since $\boldsymbol{v}=\boldsymbol{p}-\boldsymbol{q}$, this condition further means $\boldsymbol{p}=\boldsymbol{q}$ □

This proof procedure could be easily applied to type I GPK cases and we have the theorem: Theorem

Theorem 2. *Let F be a unit ball in a universal RKHS H, defined on the compact metric space X, with associated continuous kernel $k(\cdot,\cdot)$. Then type I $GPK[F,\boldsymbol{p},\boldsymbol{q}]=0$ if and only if $\boldsymbol{p}=\boldsymbol{q}$.*

Proof. Following the proof of theorem 1., we just need to add $\boldsymbol{v}=\phi(\boldsymbol{p},\boldsymbol{q})=0$ implies $\boldsymbol{p}=\boldsymbol{q}$, given the definition of type I GPK, we will have the same result as in theorem 1. □

The theorem guarantees that GPK is a family of functions with potential usage for two-sample test.

The above results may look trivial, since given the universal properties of universal-RKHS, all functions on metric space X rely in the same RKHS, so do $\boldsymbol{p},\boldsymbol{q}$ and $\phi(\boldsymbol{p},\boldsymbol{q})$. That is to say, no matter how we change the mapping function ϕ as defined in GPK, as long as the kernel function K is fixed, we are always dealing with the same RKHS. However, different mapping function ϕ will make a difference when we consider estimating these GPK directly using plugin(empirical) estimators.

4.3 PLUGIN-ESTIMATOR FOR GPK

Here we propose to estimate the term above empirically, i.e. replace $p(x),q(x)$ with its empirical frequencies

Definition 4. *We define the plugin-estimator of a given member of GPK family as*

$$\mathbb{GPK}_E[\mathcal{F},p,q]=\sum_{i\in[k]}\sum_{j\in[k]}\phi\left(\frac{N_i}{n_1}-\frac{M_i}{m_1}\right)K_{ij}\phi\left(\frac{N_j}{n_2}-\frac{M_j}{m_2}\right)$$

Where n_1, n_2, m_1, m_2 denotes different i.i.d samples drawn from distribution p and q with sample size n_1, n_2, m_1, m_2. Here our setting is different from that of Gretton et al. (2012), where in their setting n_1, n_2 represent the same sample from p and so does for m_1, m_2 and q. Another way of viewing this is that for our setting, given two samples $\{x_1,x_2,...,x_n\}$, $\{y_1,y_2,...,y_n\}$ from p and q, we depart each sample of x and y into two parts, yielding 4 different samples with size n_1, n_2, m_1, m_2, and then calculate the empirical frequencies for plugin-estimator defined above.

The reason for doing this is that it is easier to analyze the unbiasedness of the resulting plugin-estimator. However, the procedure of Gretton et al. (2012) will definitly result in higher sample efficiency and analyzing the effect of such procedure for our plugin-estimator remains future work.

The introducing of plugin-estimator is one of our reasons for proposing the GPK family. As we have already shown, all memebers in $GPK[F]$ share a common RKHS and it is not clear how

diferent mapping function ϕ will lead to the difference in measure discrepency between distributions. However, if we consider that all members in $GPK[F]$ are estimated using plugin-estimators, the difference is clear. Using different mapping function ϕ, some members of $GPK[F]$ have unbiased plugin-estimators, some members have not. Also, the variance of plugin-estimator for different members are different, yielding the arts of choosing best $GPK[F]$ member for task such as two-sample test.

We firstly illustate our techniques of analyzing bias of plugin estimator through the example of MMD case.

4.3.1 PLUGIN ESTIMATOR OF MMD IS UNBIASED

As we already know from Gretton et al. (2012), MMD has unbiased estimator building with U-statistics and since our setting of MMD as a member of GPK is actually equivalent to original MMD setting, we can decompose the plugin-estimator back into sample-based estimator as in Gretton et al. (2012):

Firstly we expand the right hand side of our original formula

$$\begin{aligned}\mathbb{MMD}_E^2[\mathcal{F},p,q] &= \sum_{i\in[k]}\sum_{j\in[k]}\left(\frac{N_i}{n_1}-\frac{M_i}{m_1}\right)K_{ij}\left(\frac{N_j}{n_2}-\frac{M_j}{m_2}\right)\\ &= \sum_{i\in[k]}\sum_{j\in[k]}\left(\frac{N_i}{n_1}k_{ij}\frac{N_j}{n_2}+\frac{M_i}{m_1}k_{ij}\frac{M_j}{m_2}-\frac{N_i}{n_1}k_{ij}\frac{M_j}{m_2}-\frac{M_i}{m_1}k_{ij}\frac{N_j}{n_2}\right)\end{aligned}$$

The right hand side is a summation of 4 parts, we label them as (1), (2), (3), (4)

We take part (1) for example, and the same procedure works for all the 4 parts.

$$\begin{aligned}\sum_{i\in[k]}\sum_{j\in[k]}\frac{N_i}{n_1}k_{ij}\frac{N_j}{n_2} &= \frac{1}{n_1n_2}\sum_{i\in[k]}\sum_{j\in[k]}N_iK_{ij}N_j\\ &= \frac{1}{n_1n_2}\sum_{x_s}^{n_1}\sum_{x_t}^{n_2}K(x_s,x_t)\end{aligned}$$

Repeat the procedure above for (1), (2), (3), (4), we get:

$$\begin{aligned}\mathrm{MMD}_E^2[\mathcal{F},X,Y] =& \frac{1}{m_1m_2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}k\left(x_i,x_j\right)+\frac{1}{n_1n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}k\left(y_i,y_j\right)\\ &-\frac{1}{m_1n_2}\sum_{i=1}^{m_1}\sum_{j=1}^{n_2}k\left(x_i,y_j\right)-\frac{1}{m_2n_1}\sum_{i=1}^{m_2}\sum_{j=1}^{n_1}k\left(x_i,y_j\right)\end{aligned}$$

Note for this procedure, x_i, x_j are from different samples, so do y_i, y_j. This estimator is the same as the unbiased estimator in Gretton et al. (2012).

Although U-statistics is a very convenient technique in building unbiased estimators, we can not always take advantage of this techniques, since for our setting of GPK family, mapping function ϕ is not always a linear function, thus GPK can not always be decomposed back into sample-based term. Here we propose the other proof procedure with the help of Bernstein polynomial that will work for all members in GPK family(see Appendix for a brief introduction of Bernstein polynomial).

Theorem 3. $\mathbb{MMD}_E^2[\mathcal{F},p,q]$ *is unbiased estimator of* $\mathbb{MMD}^2[\mathcal{F},p,q]$

Proof. Recall

$$\mathbb{MMD}_E^2[\mathcal{F},p,q] = \sum_{i\in[k]}\sum_{j\in[k]}\left(\frac{N_i}{n_1}k_{ij}\frac{N_j}{n_2}+\frac{M_i}{m_1}k_{ij}\frac{M_j}{m_2}-\frac{N_i}{n_1}k_{ij}\frac{M_j}{m_2}-\frac{M_i}{m_1}k_{ij}\frac{N_j}{n_2}\right)$$

The right hand side is a summation of 4 parts, we label them as (1), (2), (3), (4)

We take part (1) for example, and the same procedure works for all the 4 parts.

notate $f_1(x) = x$, thus the Bernstein polynomial of f_1 is $B_n(f_1, p) = p$

$$\begin{aligned}\mathbb{E}\sum_{i\in[k]}\sum_{j\in[k]}\frac{N_i}{n_1}k_{ij}\frac{N_j}{n_2} &= \sum_{i\in[k]}\mathbb{E}_{N_i\sim bin(n_1,p_i)}\left(\frac{N_i}{n_1}\sum_{j\in[k]}\mathbb{E}_{N_j\sim bin(n_2,p_j)}\frac{N_j}{n_2}K_{ij}\right)\\ &= \sum_{i\in[k]}\sum_{j\in[k]}p_iK_{ij}p_j\end{aligned}$$

We put our detailed derivation of above equation in Appendix.

Repeat the procedure above for (1), (2), (3), (4), we get:

$$\mathbb{E}\left(\mathbb{MMD}_E^2[\mathcal{F}, p, q]\right) = \mathbb{MMD}^2[\mathcal{F}, p, q]$$

Thus the empirical estimation we proposed above is an unbiased estimator of MMD.

□

5 MEMBERS IN THE GPK FAMILY AND KERNELS BUILT FROM GPK FAMILY

In section 2, we have introduced GPK family through analyzing MMD. We have already known MMD is a member of GPK family with unbiased plugin-estimator thus suitable for the two-sample test, but it is still unclear whether there exsits other members equiped with the same property.

Another interesting question is that, since members in GPK family are kernel functions, the composition rule of kernel function also apply to them. That is to say, we can build new kernel functions based on existing GPK family, the properties of the resulting kernel functions is also of interests.

5.1 KERNELIZED STEIN DISCREPANCY AND ITS PLUGIN ESTIMATOR

Our first example of GPK is the kernelized Stein discrepancy(KSD). KSD is originally proposed for goodness-of-fit, where only one sample $\{x_1, x_2, ..., x_n\} \in p$ is given, and one probability model is used to provide the evaluation of the sample on q defined by the model. The beauty of KSD is that through the introducing of Stein operator, one does not need to sample from q to evaluate the discrepancy between q and p. Although the KSD is built for goodness-of-fitting setting and has realized great success, it is still interesting to analyze it under the two-sample test, and its relationship to MMD. Our GPK framework provides a special viewpoint of this.

We firstly introduce the existing results for KSD, more specifically, the KSD for discrete distribution as proposed by Yang et al. (2018).

Definition 5 (discrete Stein discrepancy)**.** *Let $\mathcal{X}$ be a finite set. For a family $\mathcal{F}$ of functions* $\mathrm{f} : \mathcal{X}^d \to \mathbb{R}^d$*, define the discrete Stein discrepancy between two positive pmfs p, q as*

$$\mathbb{D}(q\|p) := \sup_{\mathbf{f}\in\mathcal{F}}\|\mathbb{E}_{\mathbf{x}\sim q}\left[\mathcal{T}_p\mathbf{f}(\mathbf{x})\right] - \mathbb{E}_{\mathbf{x}\sim p}\left[\mathcal{T}_p\mathbf{f}(\mathbf{x})\right]\|$$

where $\mathcal{A}_p\mathbf{f}(\mathbf{x}) = \mathbf{s}_p(\mathbf{x})\mathbf{f}(\mathbf{x})^\top - \Delta^*\mathbf{f}(\mathbf{x})$ is the different Stein operator w.r.t. $\boldsymbol{p}$. Qiang et al. (2016)

For the setting of discrete distribution, follow Yang et al. (2018):

Let $\mathcal{L}$ be any operator defined on the space of functions $\mathcal{F} = \left\{f : \mathcal{X}^d \to \mathbb{R}\right\}$ that can be written in the form 5

$$\mathcal{L}f(\mathbf{x}) = \sum_{\mathbf{x}'\in\mathcal{X}^d} g\left(\mathbf{x}, \mathbf{x}'\right) f\left(\mathbf{x}'\right), \quad \forall f \in \mathcal{F}$$

for some bivariate (possibly vector-valued) functions g on $\mathcal{X}^d \times \mathcal{X}^d$. Define a dual operator $\mathcal{L}^*$ via

$$\mathcal{L}^* f(\mathbf{x}) = \sum_{\mathbf{x}' \in \mathcal{X}^d} g(\mathbf{x}', \mathbf{x}) f(\mathbf{x}'), \quad \forall f \in \mathcal{F}$$

Theorem 4. *Note $\mathcal{F} = \left\{ f : \mathcal{X}^d \to \mathbb{R} \right\}$. For any positive probability mass function $\boldsymbol{p}$ on $\mathcal{X}^d$, a linear operator $\mathcal{T}_p$ satisfies Stein's identity*

$$\mathbb{E}_{\mathbf{x} \sim p} \left[\mathcal{T}_p f(\mathbf{x}) \right] = 0$$

for all functions $f \in \mathcal{F}$ if and only if there exist linear operators $\mathcal{L}$ and $\mathcal{L}^*$ of the forms (5) and (6), such that

$$\mathcal{T}_p f(\mathbf{x}) = \frac{\mathcal{L}p(\mathbf{x})}{p(\mathbf{x})} f(\mathbf{x}) - \mathcal{L}^* f(\mathbf{x})$$

holds for all $\mathbf{x} \in \mathcal{X}^d$ and functions $f \in \mathcal{F}$.

Theorem 5. *Denote $s_p(x) = \frac{Lp(x)}{p(x)}$ The KDSD could be written as*

$$\mathbb{D}(q\|p)^2 = \mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim q} \left[\left(s_p(x) - s_q(x) \right) k(\mathbf{x}, \mathbf{x}') \left(s_p(x') - s_q(x') \right) \right]$$

So far we have introduced all preliminary we need for analyzing KSD under view point of GPK. Next we illustrate that KDSD we defined above belongs to GPK family.

Theorem 6. *$KDSD[F, \boldsymbol{p}, \boldsymbol{q}]$ we defined above is a member of $GPK[F]$ family*

Proof. We firstly rewrite expectation into form of weighted sum

$$\begin{aligned}
\mathbb{D}(q\|p)^2 &= \mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim q} \left[\left(s_p(x) - s_q(x) \right) k(\mathbf{x}, \mathbf{x}') \left(s_p(x') - s_q(x') \right) \right] \\
&= \sum_{i \in [k]} \sum_{j \in [k]} \left(\frac{Lp_i}{p_i} - \frac{Lq_i}{q_i} \right) q_i K_{ij} \left(\frac{Lp_j}{p_j} - \frac{Lq_j}{q_j} \right) q_j
\end{aligned}$$

We define $\phi(\boldsymbol{p}, \boldsymbol{q}) = \boldsymbol{q}(L\boldsymbol{p} * /\boldsymbol{p} - L\boldsymbol{q} * /\boldsymbol{q})$ where $*/$ means element-wise division of two vectors in each dimension. We can easily see KDSD is a member of GPK family □

Note that original objective of KDSD is goodness-of-fit test, thus the ground truth for distribution $p(x)$ is already known(the distribution given by the model), thus it is safe to justing sampling according to $q(x)$ and use $p(x)$ for evaluation. However for two-sample test, $p(x)$ should be treated equally with $q(x)$. Bearing this intuition, we propose a symmetric form of KDSD for two-sample test scenario:

Definition 6. *We define symmetric-KDSD(SKDSD) to be*

$$\mathbb{SKDSD}(q\|p)^2 = \left\| \mathbb{E}_{\mathbf{x} \sim q} \left[\mathcal{T}_p \mathbf{f}(\mathbf{x}) \right] - \mathbb{E}_{\mathbf{x} \sim p} \left[\mathcal{T}_q \mathbf{f}(\mathbf{x}) \right] \right\|_{\mathcal{H}}^2$$

The SKDSD is a new probability kernel function by summation of two KDSD probability kernel, thus it is a kernel function generated by composition of members in $GPK[F]$ family, and the plugin-estimator of this probability kernel function is unbiased.

Theorem 7. *$\mathbb{SKDSD}_E^2[\mathcal{F}, p, q]$ is unbiased estimator of $\mathbb{SKDSD}^2[\mathcal{F}, p, q]$*

Proof.

$$\begin{aligned}
\mathbb{SKDSD}(q\|p)^2 &= \left\| \mathbb{E}_{\mathbf{x} \sim q} \left[\mathcal{T}_p \mathbf{f}(\mathbf{x}) \right] - \mathbb{E}_{\mathbf{x} \sim p} \left[\mathcal{T}_q \mathbf{f}(\mathbf{x}) \right] \right\|_{\mathcal{H}}^2 \\
&= \left\| \sum_{i \in [k]} \mathcal{T}_{p_i} \mathbf{f}(\mathbf{x_i}) q_i - \sum_{i \in [k]} \mathcal{T}_{q_i} \mathbf{f}(\mathbf{x_i}) p_i \right\|_{\mathcal{H}}^2 \\
&= \left\| \sum_{i \in [k]} \left(\frac{Lp_i}{p_i} \mathbf{f}(\mathbf{x_i}) - L^* \mathbf{f}(\mathbf{x_i}) \right) q_i - \sum_{i \in [k]} \left(\frac{Lq_i}{q_i} \mathbf{f}(\mathbf{x_i}) - L^* \mathbf{f}(\mathbf{x_i}) \right) p_i \right\|_{\mathcal{H}}^2 \\
&= \sum_{i \in [k]} \sum_{j \in [k]} Lp_i Lq_j K_{ij} - Lp_i q_j K_{ij}^* - Lq_i p_j K_{ij}^* + p_i q_j K_{ij}^{**}
\end{aligned}$$

since the operator L, L^* is a linear operator, the same proof procedure of MMD also works for SKDSD □

5.2 POLYNOMIAL PROBABILITY KERNEL

Next we analyze a simple group of members belongs to GPK family

Definition 7 (polynomial probability kernel)**.** *A polynomial probability kernel* $\mathbb{PO}_{lk}[F, \boldsymbol{p}, \boldsymbol{q}]$ *is a special case of type I GPK which mapping function* ϕ *is a polynomial function* $\phi(\boldsymbol{p}, \boldsymbol{q}) = \boldsymbol{p}^l - \boldsymbol{q}^k$

As we can see, MMD is a member of polynomial probability kernel as $\mathbb{PO}_{11}[F, \boldsymbol{p}, \boldsymbol{q}]$

As a brief example, we illustrate the polynomial probability kernel with mapping function $\phi(\boldsymbol{p}, \boldsymbol{q}) = \boldsymbol{p}^2 - \boldsymbol{q}^2$. Interestingly, we show that although its plugin-estimator is biased, it could be modified to be unbiased. This result is obtained through analyzing the unbiasedness of the plugin-estimator of this GPK family which illustrates the power of our framework.

Theorem 8. *the default plugin-estimator of* $\mathbb{PO}_{22}[\mathcal{F}, \boldsymbol{p}, \boldsymbol{q}]$ *is a biased estimator*

Proof. We put our formal proof in Appendix A.3 □

The interesting result is that by including $\mathbb{PO}_{12}[\mathcal{F}, \boldsymbol{p}, \boldsymbol{q}]$, $\mathbb{PO}_{21}[\mathcal{F}, \boldsymbol{p}, \boldsymbol{q}]$ and $\mathbb{PO}_{11}[\mathcal{F}, \boldsymbol{p}, \boldsymbol{q}]$ into consideration, we can see the linear combination the default of empirical-estimators of these probability kernels may yield unbiased estimators. The basic idea is to check the bias of the each $\mathbb{PO}^2_{lkE}[\mathcal{F}, p, q]$ using the techniques of Bernstein polynomial, and then correct the bias with another $\mathbb{PO}^2_{lkE}[\mathcal{F}, p, q]$.

Theorem 9. $\frac{m}{m-1}\left(\mathbb{PO}^2_{12E}[\mathcal{F}, p, q] - \frac{1}{m}\mathbb{PO}^2_{1E}[\mathcal{F}, p, q]\right)$ *is an unbiased estimator of* $\mathbb{PO}^2_{12}[\mathcal{F}, p, q]$

Proof. We put our detailed proof in appendix A.4 □

Theorem 10. $\left(\frac{m}{m-1}\right)^2\left(\mathbb{PO}^2_{22E}[\mathcal{F}, p, q] - \frac{1}{m}\mathbb{PO}^2_{12E}[\mathcal{F}, p, q] - \frac{1}{m}\mathbb{PO}^2_{21E}[\mathcal{F}, p, q] + \frac{1}{m^2}\mathbb{PO}^2_{12E}[\mathcal{F}, p, q]\right)$ *is an unbiased estimator of* $\mathbb{PO}^2_{22}[\mathcal{F}, p, q]$

Proof. We put our detailed proof in appendix A.5 □

6 TWO-SAMPLE TEST USING GENERALIZED PROBABILITY KERNEL

In this section, we briefly discuss the potential usage of the GPK family in two-sample test. We firstly summarize the requirements for a statistic to be suitable for the two-sample test, then come to each of the special GPK family members we have discussed so far.

Following the work of Gretton et al. (2012), we summarize that, for a statistic to be suitable for the two-sample test, it should:

- provide the property that $D(p\|q) = 0$ if and only if $p = q$
- should have unbiased estimators, or at least estimators with bounded bias
- the estimator in requirement 2. should have the variance is analyzable so as to provide valid test threshold and test power.

The first requirement is satisfied by all the members of type I GPK we have defined, and this includes MMD, $\mathbb{PO}_{22}[F, \boldsymbol{p}, \boldsymbol{q}]$ as examples. The second requirement is satisfied by the members of GPK family with unbiased plugin-estimators, which includes MMD, SKDSD $\mathbb{PO}_{22}[F, \boldsymbol{p}, \boldsymbol{q}]$ as examples.

The third requirement has not been discussed in this paper so far, but using the technique of Bernstein polynomial anlysis, it is possible and remain further study.

REFERENCES

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 12:723–773, 2012.

Tony Jebara, Risi Kondor, and Andrew Howard. Probability product kernels. *J. Mach. Learn. Res.*, 5:819–844, December 2004. ISSN 1532-4435.

Liu Qiang, Lee Jason, and Jordan Michael. A kernelized stein discrepancy for goodness-of-fit tests. *International Conference on Machine Learning*, 48:276–284, 2016.

Jiasen Yang, Qiang Liu, Vinayak Rao, and Jennifer Neville. Goodness-of-fit testing for discrete distributions via stein discrepancy. *International Conference on Machine Learning*, 80:276–284, 2018.

Hao Yi and Orlitsky Alon. Data amplification: Instance-optimal property estimation. *International Conference on Machine Learning*, 2020.

A Appendix

A.1 Bernstein polynomial

Drawing i.i.d. samples Y^m from any distribution p, the expected value of the empirical estimator for a distribution property is

$$\mathbb{E}\left[\hat{H}^{\mathrm{E}}\left(Y^m\right)\right] = \sum_{i\in[k]} \mathop{\mathbb{E}}_{M_i\sim \mathrm{bin}(m,p_i)}\left[h\left(\frac{M_i}{m}\right)\right]$$

Note that for any function $f, m \in \mathbb{N}$, and $x \in [0,1]$, the degree-m Bernstein polynomial of f is

$$B_m(f,x) := \sum_{j=0}^{m} f\left(\frac{j}{m}\right)\left(\begin{array}{c} m \\ j \end{array}\right) x^j (1-x)^{m-j}$$

Therefore, we can express the expectation of the empirical property estimator as

$$\mathop{\mathbb{E}}_{Y^m\sim p}\left[\hat{H}^{\mathrm{E}}\left(Y^m\right)\right] = \sum_{i\in[k]} B_m\left(h, p_i\right)$$

A.2 Detailed proof of Theorem 1.

Proof.

$$\begin{aligned}\mathbb{MMD}_E^2[\mathcal{F},p,q] &= \sum_{i\in[k]}\sum_{j\in[k]}\left(\frac{N_i}{n}-\frac{M_i}{m}\right)K_{ij}\left(\frac{N_j}{n}-\frac{M_j}{m}\right)\\ &= \sum_{i\in[k]}\sum_{j\in[k]}\left(\frac{N_i}{n}k_{ij}\frac{N_j}{n}+\frac{M_i}{m}k_{ij}\frac{M_j}{m}-\frac{N_i}{n}k_{ij}\frac{M_j}{m}-\frac{M_i}{m}k_{ij}\frac{N_j}{n}\right)\end{aligned}$$

The right hand side is a summation of 4 parts, we label them as (1), (2), (3), (4)

We take part (1) for example, and the same procedure works for all the 4 parts.

notate $f_1(x) = x$, thus the Bernstein polynomial of f_1 is $B_n(f_1, p) = p$

$$
\begin{aligned}
\mathbb{E}\sum_{i\in[k]}\sum_{j\in[k]}\frac{N_i}{n}k_{ij}\frac{N_j}{n} &= \sum_{i\in[k]}\mathbb{E}_{N_i\sim bin(n,p_i)}\left(\frac{N_i}{n}\sum_{j\in[k]}\mathbb{E}_{N_j\sim bin(n,p_j)}\frac{N_j}{n}K_{ij}\right)\\
&= \sum_{i\in[k]}\mathbb{E}_{N_i\sim bin(n,p_i)}\left(\frac{N_i}{n}\sum_{j\in[k]}K_{ij}B_n(f_1,p_j)\right)\\
&= \sum_{i\in[k]}\mathbb{E}_{N_i\sim bin(n,p_i)}\left(\frac{N_i}{n}\sum_{j\in[k]}K_{ij}p_j\right)\\
&= \sum_{i\in[k]}\left(\sum_{j\in[k]}K_{ij}p_j\right)B_n(f_1,p_i)\\
&= \sum_{i\in[k]}\sum_{j\in[k]}p_iK_{ij}p_j
\end{aligned}
$$

Repeat the procedure above for (1), (2), (3), (4), we get:

$$
\mathbb{E}\left(\mathbb{MMD}_E^2[\mathcal{F},p,q]\right) = \mathbb{MMD}^2[\mathcal{F},p,q]
$$

Thus the empirical estimation we proposed above is an unbiased estimator of MMD

□

A.3 Detailed proof of Theorem 8.

Proof. We begin with $\mathbb{PO}_{12E}^2[\mathcal{F},p,q]$ and the result also works for $\mathbb{PO}_{21E}^2[\mathcal{F},p,q]$

$$
\begin{aligned}
\mathbb{PO}_{12E}^2[\mathcal{F},p,q] &= \sum_{i\in[k]}\sum_{j\in[k]}f_{lk}(\frac{N_i}{n},\frac{M_i}{m})K_{ij}f_{lk}(\frac{N_j}{n},\frac{M_j}{m})\\
&= \sum_{i\in[k]}\sum_{j\in[k]}\left(\frac{N_i}{n}-\frac{M_i}{m}^2\right)K_{ij}\left(\frac{N_j}{n}-\frac{M_j}{m}^2\right)\\
&= \sum_{i\in[k]}\sum_{j\in[k]}\left(\frac{N_i}{n}\frac{N_j}{n}+\frac{M_i}{m}^2\frac{M_j}{m}^2-\frac{N_i}{n}\frac{M_j}{m}^2-\frac{M_i}{m}^2\frac{N_j}{n}\right)K_{ij}
\end{aligned}
$$

this estimator of first part is unbiased, but not for the remaining parts, let's start from the third part

$$
\begin{aligned}
\mathbb{E}\left[\mathbb{PO}^2_{12E}[\mathcal{F},p,q]_{part3}\right] &= \sum_{i\in[k]}\sum_{j\in[k]} \mathbb{E}_{N_i\sim bin(n,p_i)}\mathbb{E}_{M_j\sim bin(m,q_j)} \frac{N_i}{n}\frac{M_j}{m}^2 K_{ij} \\
&= \sum_{j\in[k]} \mathbb{E}_{M_j\sim bin(m,q_j)} \frac{M_j}{m}^2 \left(\sum_{i\in[k]} \mathbb{E}_{N_i\sim bin(n,p_i)} \frac{N_i}{n} K_{ij}\right) \\
&= \sum_{j\in[k]} \mathbb{E}_{M_j\sim bin(m,q_j)} \frac{M_j}{m}^2 \left(\sum_{i\in[k]} b_n(f_1,p_i) K_{ij}\right) \\
&= \sum_{j\in[k]} \mathbb{E}_{M_j\sim bin(m,q_j)} \frac{M_j}{m}^2 \left(\sum_{i\in[k]} p_i K_{ij}\right) \\
&= \sum_{j\in[k]} b_m(f_2,q_j) \left(\sum_{i\in[k]} p_i K_{ij}\right) \\
&= \sum_{j\in[k]} \left(\frac{m-1}{m}q_j^2 + \frac{q_j}{m}\right) \left(\sum_{i\in[k]} p_i K_{ij}\right) \\
&\neq \sum_{i\in[k]}\sum_{j\in[k]} q_j^2 p_i K_{ij} = \mathbb{PO}^2_{12}[\mathcal{F},p,q]_{part3}
\end{aligned}
$$

the same procedure follows for part2 and part4, so as to $\mathbb{PO}^2_{22E}[\mathcal{F},p,q]$ □

A.4 Detailed proof of Theorem 9.

Proof. back to the part3 of $\mathbb{PO}^2_{12E}[\mathcal{F},p,q]$, we have

$$
\begin{aligned}
\mathbb{E}\left[\mathbb{PO}^2_{12E}[\mathcal{F},p,q]_{part3}\right] &= \sum_{j\in[k]} \left(\frac{m-1}{m}q_j^2 + \frac{q_j}{m}\right) \left(\sum_{i\in[k]} p_i K_{ij}\right) \\
&= \frac{m-1}{m}\sum_{j\in[k]}\sum_{i\in[k]} q_j^2 p_i K_{ij} + \frac{1}{m}\sum_{j\in[k]}\sum_{i\in[k]} q_j p_i K_{ij}
\end{aligned}
$$

thus

$$
\begin{aligned}
\mathbb{E}\left[\frac{m}{m-1}\left(\mathbb{PO}^2_{12E}[\mathcal{F},p,q]_{part3} - \frac{1}{m}\mathbb{PO}^2_{1E}[\mathcal{F},p,q]_{part3}\right)\right] &= \sum_{i\in[k]}\sum_{j\in[k]} q_j^2 p_i K_{ij} \\
&= \mathbb{PO}^2_{12}[\mathcal{F},p,q]_{part3}
\end{aligned}
$$

□

A.5 Detailed proof of Theorem 10.

Proof. back to the part4 of $\mathbb{PO}^2_{12E}[\mathcal{F},p,q]$, we have

$$
\begin{aligned}
\mathbb{E}\left[\mathbb{PO}^2_{12E}[\mathcal{F},p,q]_{part4}\right] &= \sum_{j\in[k]} \left(\frac{m-1}{m}q_j^2 + \frac{q_j}{m}\right) \left(\sum_{i\in[k]} \left(\frac{m-1}{m}p_i^2 + \frac{p_i}{m}\right) K_{ij}\right) \\
&= \left(\frac{m-1}{m}\right)^2 \sum_{j\in[k]}\sum_{i\in[k]} q_j^2 p_i^2 K_{ij} + \frac{1}{m}\mathbb{E}\left[\mathbb{PO}^2_{12E}[\mathcal{F},p,q]\right] \\
&+ \frac{1}{m}\mathbb{E}\left[\mathbb{PO}^2_{21E}[\mathcal{F},p,q]\right] - \frac{1}{m^2}\mathbb{E}\left[\mathbb{PO}^2_{1E}[\mathcal{F},p,q]\right]
\end{aligned}
$$

thus

$$\begin{aligned}\mathbb{E}PO =& \left(\frac{m}{m-1}\right)^2 \left(\mathbb{PO}^2_{22E}[\mathcal{F},p,q] - \frac{1}{m}\mathbb{PO}^2_{12E}[\mathcal{F},p,q] - \frac{1}{m}\mathbb{PO}^2_{21E}[\mathcal{F},p,q] + \frac{1}{m^2}\mathbb{PO}^2_{12E}[\mathcal{F},p,q]\right) \\ =& \mathrm{PO}^2_{22}[\mathcal{F},p,q]\end{aligned}$$

□