Ranking Function Based Parameter Estimation

Benedek B. Novák
Supervisor: Balázs Csanád Csáji

I. Introduction

The aim of this report is to discuss and elaborate on the generalization of the Ranking function introduced by Ádám Jung and Balázs Csanád Csáji [1], who used it to make predictions based on the resampling framework introduced by Ambrus Tamás and Balázs Csanád Csáji [2].

The advantage of the discussed methods lies in their generality. Our framework will be the following: We have an i.i.d. generated sample from $\mathbb{P}_{\vartheta^*}$ parameterized by $\vartheta^* \in \Theta$ where Θ is a metric space. Furthermore, we have access to a black box, that can generate samples based on a given parameter - or in a bit less general, but more preferable case, it can reproduce the distribution $\mathbb{P}_\vartheta$ using a sample from a uniform distribution.

The main goal is to find the parameter $\hat{\vartheta}$ that generates the most similar data to the original one. The way to construct these similarity measures has a great flexibility, but in this report, we will discuss MMD distance based approaches, which also have the benefit of being applicable in very general cases and greatly customisable with the choice of the kernel function. Finally, the theoretical support for our estimation will come from the resampling framework, which we will present shortly.

II. Kernel Mean Embeddings and MMD

For the sake of completeness, we begin our discussion with a brief introduction on Reproducing Kernel Hilbert Spaces and Kernel Mean Embeddings, since they will be a useful tool to compare the similarity of probability distributions later.

Definition II.1. [3] Let $\mathcal{X}$ be an arbitrary set, and $\mathcal{H}$ a Hilbert space of $\mathcal{X} \to \mathbb{R}$ functions and denote the *evaluation functional* with $E_x : \mathcal{H} \to \mathbb{R}$ (i.e. $E_x(f) = f(x)$). $\mathcal{H}$ is called a *Reproducing Kernel Hilbert Space* (RKHS), if all of its evaluation functionals are bounded is the sense that there exists a $C_x > 0$ for all E_x such that $|E_x(f)| \le C_x ||f||_{\mathcal{H}}$.

From the Riesz representation theorem it follows that for all $x \in \mathcal{X}$ there exists a $k_x \in \mathcal{H}$ s.t. $f(x) = \langle f, k_x \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$.

Definition II.2. [3] The *reproducing kernel* of RKHS $\mathcal{H}$ over $\mathcal{X}$ is the function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined as $k(x, y) := \langle k_y, k_x \rangle_{\mathcal{H}}$.

Remark. From the statements above it follows that $k_x = k(\cdot, x)$, therefore $f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$ and $x \in \mathcal{X}$. We call this the reproducing property.

Definition II.3. [3] We say that a kernel function k is *positive definite*, if for any finite $\{x_1, ..., x_n\} \subseteq \mathcal{X}$ and $\{a_i\}_{i=1}^n \subset \mathbb{R}$ it holds that

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j k(x_i, x_j) \ge 0$$

Definition II.4. [4] The kernel mean embedding of a probability measure $\mathbb{P}$ into an RKHS $\mathcal{H}$ is defined as

$$\mu_{\mathbb{P}} = \int k(\cdot, x) \, \mathrm{d}\mathbb{P}(x)$$

Here the integral is to be interpreted as a Bochner-integral, as defined in [4] in a similar manner to the Lebesgue integral.

Remark. If $E_{X\sim\mathbb{P}}[\sqrt{k(X,X)}] < \infty$, then $\mu_{\mathbb{P}} \in \mathcal{H}$ and $E_{X\sim\mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$.

Definition II.5. [4] The *Maximum Mean Discrepancy* (MMD) of two distributions, $\mathbb{P}$ and $\mathbb{Q}$ is defined as the distance of their mean embeddings in the RKHS:

$$\mathrm{MMD}^2_{\mathcal{H}}[\mathbb{P}, \mathbb{Q}] = ||\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}||^2_{\mathcal{H}}$$

The MMD can be estimated with an unbiased estimator using samples X, Y from the distributions $\mathbb{P}, \mathbb{Q}$ with sizes n and m respectively:

$$\begin{aligned}\widehat{\mathrm{MMD}}^2_{\mathcal{H}}[X, Y] =& \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1}^n k(x_i, x_j) \\ &+ \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j=1}^m k(y_i, y_j) \\ &- \frac{2}{nm} \sum_{i=1}^m \sum_{j=1}^n k(x_i, y_j)\end{aligned}$$

III. The Resampling Framework

Now we proceed by introducing the hypothesis tests that will be used to provide the theoretical support for methods and the estimated parameters. In order to construct these hypothesis tests, let $\mathcal{P} = \{\mathbb{P}_\vartheta \in \Theta\}$ be class of non-atomic (i.e. $\mathbb{P}_\vartheta(x) = 0 \ \forall x \in \mathcal{X}$) probability distributions over $\mathcal{X} \subseteq \mathbb{R}^d$, where Θ is the parameter space. We also assume that there is a distribution $\mathbb{P}_{\vartheta^*} \in \mathcal{P}$, from which we receive an i.i.d. sample $\{x_1, ..., x_n\} \subseteq \mathcal{X}$. The goal is to construct hypothesis tests for $H_0 : \mathbb{P}_\vartheta = \mathbb{P}_{\vartheta^*}$ and $H_1 : \mathbb{P}_\vartheta \neq \mathbb{P}_{\vartheta^*}$. The core idea of the framework is to generate $m-1$ i.i.d. sets of alternate samples with n elements, each from $\mathbb{P}_\vartheta$ in order to perform the hypothesis test. We denote the original sample with $\mathcal{S}^{(0)}$, and the i-th alternative sample with $\mathcal{S}^{(i)}(\vartheta) = (x_1^{(i)}(\vartheta), ..., x_n^{(i)}(\vartheta))$.

Remark. There are two distinct cases to consider when generating the alternative samples.
In the first case, we have access to a black box, that given a ϑ parameter, generates i.i.d. samples from P_ϑ at random. In the second case, again, we have a black box B, but this time the black box itself doesn't contain any randomness: it is a function that gives the same output every time for every $(\vartheta, q) \in \Theta \times [0,1]^d$ pair. However, if q is drawn from the uniform distribution $U[0,1]^d$, then $B(\vartheta, q)$ has distribution $\mathbb{P}_\vartheta$. Examples for the second case are the inverses of the cumulative distribution functions, or neural networks that given only random noise can generate meaningful samples. (For example diffusion models for image generation.)

Note that in the second case, $\mathcal{S}^{(i)}(\vartheta)$ can be thought of as a a function of ϑ if we fix a sample from the uniform distribution a priori.

Definition III.1. [2] Let $\mathbb{A}$ be a measurable space, denote $\{1, ..., m\}$ with $[m]$. Then $\psi : \mathbb{A}^m \to [m]$ is a ranking function if it satisfies the following properties:
P1) Invariance with regards to the reordering of the last $m-1$ elements, i.e. for all $(a_1, ..., a_m) \in \mathbb{A}^m$:

$$\psi(a_1, a_2, ..., a_m) = \psi(a_1, a_{\pi(2)}, ..., a_{\pi(m)})$$

where π is a permutation on the set $\{2, ..., m\}$.
P2) Uniqueness in the first variable, i.e. for all $i, j \in [m]$ if $a_i \neq a_j$, then

$$\psi(a_i, \{a_k\}_{k\neq i}) \neq \psi(a_j, \{a_k\}_{k\neq j})$$

where the shorthand notation is justified by P1.

Ranking functions are mostly based on reference variables denoted by:

$$Z^{(i)}(\vartheta) := T(\mathcal{S}^{(i)}(\vartheta), \vartheta)$$

where $T : \mathcal{X}^n \times \Theta \to \mathbb{R}$. ($Z^{(0)}(\vartheta)$ is defined as $T(S^{(0)}, \vartheta)$.) These reference variables might seem a bit arbitrary at first, so let's have a look at some examples:

Example. The maximum likelihood based reference variables: If $\mathcal{L}(\vartheta, S^{(i)})$ denotes the log-likelihood of sample $S^{(i)}$, then

$$Z^{(i)}(\vartheta) = ||\nabla_\vartheta \mathcal{L}(\vartheta, S^{(i)})||^2$$

Unfortunately to use these as reference variables, we need the derivative of the log-likelihood function, which makes them generalize poorly.

Example. MMD based reference variables (with an arbitrary choice of $\mathcal{H}$ RKHS):
The first construction is based on the similarity of the sample to all other samples:

$$Z^{(i)}(\vartheta) = \sum_{j=0}^{n-1} \widehat{\text{MMD}}^2_{\mathcal{H}}[S^{(i)}(\vartheta), S^{(j)}(\vartheta)]$$

This has the benefit of not needing any further a priori knowledge about the distributions, other than that they are parameterized by a $\vartheta \in \Theta$. However, since we need to compare every sample to every other sample, the runtime becomes quadratic in the number of resamplings m, so to combat this, another unbiased estimator can be constructed by doing one extra resampling from the distribution, denoted by $S^{(m)}(\vartheta)$:

$$Z^{(i)}(\vartheta) = \widehat{\text{MMD}}^2_{\mathcal{H}}[S^{(i)}(\vartheta), S^{(m)}(\vartheta)]$$

Here, we compare all samples $S^{(0)}, S^{(1)}(\vartheta), ..., S^{(m-1)(\vartheta)}$ to $S^{(m)}(\vartheta)$ to obtain $Z^{(0)}(\vartheta), ..., Z^{(m-1)}(\vartheta)$

These Reference variables are then sorted in ascending order, so the rank of $S^{(i)}(\vartheta)$ becomes its its place in the ordering, i.e.

$$\psi(S^{(i)}(\vartheta), \{S^{(j)}(\vartheta)\}_{j\neq i}) = 1 + \sum_{j\neq i} \mathbb{I}_{\{Z^{(j)}(\vartheta) < Z^{(i)}(\vartheta)\}}$$

Unfortunately, these reference variables could sometimes take on the same values for some ϑ, so to insure a strict ordering, a pseudo-ordering can included in the ranking function:

Definition III.2. [2] Let $\pi : [m] \to [m]$ be a random permutation, which we select random uniformly from the set of all such permutations. Then we say that $Z^{(i)}(\vartheta) <_\pi Z^{(j)}(\vartheta)$ if $Z^{(i)}(\vartheta) < Z^{(j)}(\vartheta)$ or $Z^{(i)}(\vartheta) = Z^{(j)}(\vartheta)$ and $\pi(i) < \pi(j)$.

With this ordering, we can ensure that the reference variable based ranking functions will indeed be ranking functions.

If $Z^{(j)}$ are constructed in such a way that a better fit between the sample and $\mathbb{P}_\vartheta$ corresponds to a lower value, then it is clear that having a lower rank on the original sample would imply a better estimate of the parameter.

Theorem III.3. *[5] Given a ranking function ψ, a parameter set Θ, and integer hyperparameters (q, m) with $1 \leq q \leq m$, under the null hypothesis $H_0 : \mathbb{P}_\vartheta = \mathbb{P}_{\vartheta^*}$ a confidence region for ϑ^* can be constructed as:*

$$\tilde{\Theta}^\psi_{(q,m)} := \{\vartheta \in \Theta | 1 \leq \psi(\mathcal{S}^{(0)}, \{\mathcal{S}^{(k)}(\vartheta)\}_{k\neq 0}) \leq q\}$$

where we have

$$\mathbb{P}(\vartheta^* \in \Theta^\psi_{(q,m)}) = \frac{q}{m}$$

IV. Parameter Estimation

From now on, we denote the rank of the original sample with regards to the $m-1$ i.i.d samples generated from $\mathbb{P}_\vartheta$ with $\mathcal{R}(\vartheta) := \psi(S^{(0)}, \{S^{(j)}(\vartheta)\}_{j\in[1,m-1]})$ Or, in terms of reference variables:

$$\mathcal{R}(\vartheta) = 1 + \sum_{j=1}^{m-1} \mathbb{I}_{\{Z^{(j)}(\vartheta) < Z^{(0)}(\vartheta)\}}$$

The key idea is that a ϑ that minimizes $\mathcal{R}$ will be a good estimation, therefore the point estimate is defined as $\hat{\vartheta} \in \underset{\vartheta\in\Theta}{\text{argmin}}\, \mathcal{R}(\vartheta)$. However, since $\mathcal{R}$ is a piecewise constant function in Θ, it is generally a hard problem to find the estimate. To combat this, the concept of smoothed rank was introduced by Ádám Jung and Balázs Csanád Csáji [1]:

$$\tilde{\mathcal{R}}(\vartheta) = \begin{cases} \dfrac{Z^{(0)}}{Z^{(1)}_*} & \text{if } Z^{(0)} < Z^{(1)}_* \\[2ex] k + \dfrac{Z^{(0)} - Z^{(k)}_*}{Z^{(k+1)}_* - Z^{(k)}_*} & \text{if } Z^{(k)}_* \leq Z^{(0)} < Z^{(k+1)}_* \\[2ex] m - 1 + \dfrac{Z^{(0)}}{Z^{(m-1)}_*} & \text{if } Z^{(m-1)}_* \leq Z^{(0)} \end{cases}$$

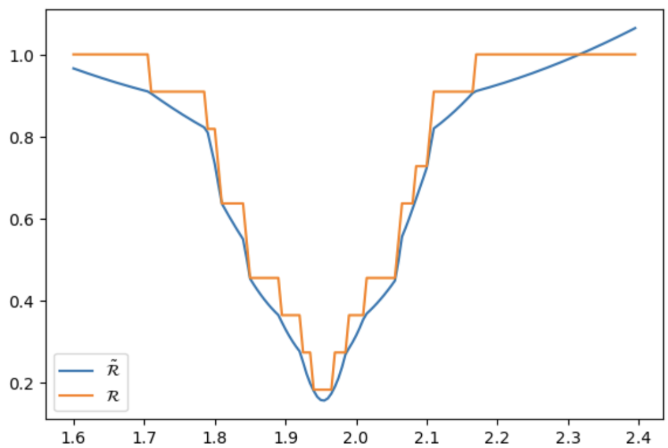

Fig. 1. relative rank ($\mathcal{R}/m$) and smoothed rank of a sample from an exponential distribution with parameter 2 using MMD based reference variables using a degree 2 polynomial kernel ($n = 250$, $m = 10$)

where $Z_*^{(1)}(\vartheta) \leq Z_*^{(2)}(\vartheta) \leq ... \leq Z_*^{(1)}(\vartheta)$ denotes the ordered version of the reference variables in each point $\vartheta \in \Theta$ and for the sake of visual clarity, the arguments (ϑ) of the reference variables were omitted. Here, we assume that $Z^{(i)}(\vartheta) \neq Z^{(j)}(\vartheta)\ \forall i \neq j$ $\mathbb{P}_\vartheta$-almost surely. Another possibility to alter $\mathcal{R}$ in order to solve the problem of optimizing a stepwise constant function without the previous assumption is with the choice of

$$\bar{\mathcal{R}}(\vartheta) = Z^{(0)}(\vartheta) - Z_*^{(1)}(\vartheta)$$

It is easy to see that even though the minima of the two functions might be at two different parameters, they will both be located in the confidence region $\Theta^{\psi}_{(1,m)}$, if it is not the empty set.

The continuity of both constructions are entirely dependent on the continuity of the reference variables, as we will show next.

Lemma IV.1. *Let (Θ, d) be a metric space, and $Z^{(i)} : \Theta \to \mathbb{R}$ $(i \in [m])$ continuous functions. Denote their ordered version with $Z_*^{(i)}$:*

$$Z_*^{(i)}(\vartheta) = \min_{j\in[m]} \left\{ Z^{(j)}(\vartheta) | \# \left\{ k | Z^{(j)}(\vartheta) \geq Z^{(k)}(\vartheta) \right\} \geq i \right\}$$

i.e. $Z_*^{(1)}(\vartheta) \leq ... \leq Z_*^{(1)}(\vartheta)$. (# denotes the cardinality of the set.) Then $Z_*^{(i)}$ are continuous for all $i \in [m]$ in Θ.

Proof. Let $B(\vartheta, \delta) = \{\vartheta' \in \Theta | d(\vartheta, \vartheta') < \delta\}$ denote the δ neighborhood of ϑ. We need to prove that for any $i \in [m], \vartheta \in \Theta$ and $\varepsilon > 0$ there exists a $\delta > 0$ such that for any $\vartheta' \in B(\vartheta, \delta)$ it holds that $|Z_*^{(i)}(\vartheta') - Z_*^{(i)}(\vartheta)| < \varepsilon$. (Note that it is enough to prove that this is true for all sufficiently small ε.)

Let $I^{(i)}(\vartheta) := \{k | Z^{(k)}(\vartheta) = Z_*^{(i)}(\vartheta)\}$ for any fixed i. Since all $Z^{(k)}$ are continuous, for sufficiently small $\varepsilon' > 0$ it holds that $I^{(i)}(\vartheta) = I_{\varepsilon'}^{(i)}(\vartheta)$ where

$$I_\varepsilon^{(i)}(\vartheta) = \left\{ k | \exists \vartheta' \in B(\vartheta, \varepsilon) : Z^{(k)}(\vartheta') = Z_*^{(i)}(\vartheta') \right\}$$

Therefore, since all $Z^{(k)}$ are continuous, there exists a δ_k for all $\varepsilon' > \varepsilon > 0$ and $k \in I(\vartheta)$ such that $\forall \vartheta' \in B(\vartheta, \delta_k) : |Z^{(k)}(\vartheta) - Z^{(k)}(\vartheta')| < \varepsilon$. Let $\delta := \min_{k \in I(\vartheta)} \{\delta_k\}$. Now, since $|Z^{(k)}(\vartheta) - Z^{(k)}(\vartheta')| < \varepsilon$ holds for all $\vartheta' \in B(\vartheta, \delta)$ and $Z_*^{(i)}(\vartheta')$ takes its value from $\{Z_*^{(k)}(\vartheta') | k \in I(\vartheta)\}$ for sufficiently small $\varepsilon > 0$, we have $|Z_*^{(i)}(\vartheta') - Z_*^{(i)}(\vartheta)| < \varepsilon$. ☐

Corollary IV.2. *If $Z^{(i)}$ are continuous in Θ, then $\bar{\mathcal{R}}(\vartheta)$ is continuous in Θ. Furthermore, if $\mathbb{P}_\vartheta(Z^{(i)}(\vartheta) = Z^{(j)}(\vartheta))$ for all $\vartheta \in \Theta$ and $i \neq j$, then $\tilde{\mathcal{R}}(\vartheta)$ is continuous with probability one.*

Proof. $\bar{\mathcal{R}}$ and $\tilde{\mathcal{R}}$ were both constructed from elementary operations of $Z^{(i)}$ and $Z_*^{(i)}$, both of which are continuous. ☐

Now that the continuity of $\tilde{\mathcal{R}}$ and $\bar{\mathcal{R}}$ is ensured, stepwise optimization techniques can be used to find their minimum.

V. Asymptotic behavior

An interesting question that can be asked is what happens if we increase the number of elements in each sample (n) or the number of subsamplings (m). From now on, $\mathcal{R}(\vartheta)$ (and $\tilde{\mathcal{R}}(\vartheta)$) will denote the *relative rank*, which is the rank divided by m, staying in the $[0, 1]$ interval for every m, so we can compare its values for different ms.

First, we investigate the asymptotic behavior for $n \to \infty$, but for this we need the following lemma:

Lemma V.1. *Law of large numbers for kernel mean embeddings: Let $\mathcal{H}$ be a real RKHS over $\mathcal{X}$ and $(\mathcal{X}, \mathcal{A}, \mathbb{P}_\vartheta)$ a probability space. Denote the empirical distribution of a sample $S(\vartheta) = \{x_1, ..., x_n\}$ with $Q_{n,\vartheta}(A) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{x_i \in A}$ for all $A \in \mathcal{A}$. If $E_{X\sim\mathbb{P}_\vartheta}[h(X)] < \infty$ for all $h \in \mathcal{H}$, then it holds that $||\mu_{\mathbb{P}_\vartheta} - \mu_{Q_{n,\vartheta}}||^2_{\mathcal{H}} \to 0$ if $n \to \infty$.*

Proof.

$$\begin{aligned}
||\mu_{\mathbb{P}_\vartheta} - \mu_{Q_{n,\vartheta}}||^2_{\mathcal{H}} &= \left\langle \mu_{\mathbb{P}_\vartheta} - \mu_{Q_{n,\vartheta}}, \mu_{\mathbb{P}_\vartheta} - \mu_{Q_{n,\vartheta}} \right\rangle_{\mathcal{H}} = \\
&= \left\langle \mu_{\mathbb{P}_\vartheta}, \mu_{\mathbb{P}_\vartheta} \right\rangle - 2\left\langle \mu_{\mathbb{P}_\vartheta}, \mu_{Q_{n,\vartheta}} \right\rangle + \left\langle \mu_{Q_{n,\vartheta}}, \mu_{Q_{n,\vartheta}} \right\rangle = \\
&= (\left\langle \mu_{Q_{n,\vartheta}}, \mu_{Q_{n,\vartheta}} \right\rangle - \left\langle \mu_{\mathbb{P}_\vartheta}, \mu_{Q_{n,\vartheta}} \right\rangle)^{(1)} + \\
&\quad + (\left\langle \mu_{\mathbb{P}_\vartheta}, \mu_{\mathbb{P}_\vartheta} \right\rangle - \left\langle \mu_{\mathbb{P}_\vartheta}, \mu_{Q_{n,\vartheta}} \right\rangle)^{(2)}
\end{aligned}$$

We will show that both (1) and (2) tend to zero as $n \to \infty$. From the definition of the kernel mean embeddings and the reproducing property of the RKHS we have:

$$\left\langle \mu_{Q_{n,\vartheta}}, \mu_{Q_{n,\vartheta}} \right\rangle_{\mathcal{H}} = \left\langle \frac{1}{n}\sum_{i=1}^n k(\cdot, x_i), \frac{1}{n}\sum_{j=1}^n k(\cdot, x_j) \right\rangle_{\mathcal{H}} =$$

$$= \frac{1}{n}\sum_{i=1}^n \frac{1}{n}\sum_{j=1}^n k(x_j, x_i) = \frac{1}{n}\sum_{i=1}^n \frac{1}{n}\sum_{j=1}^n k_{x_i}(x_j)$$

$$\left\langle \mu_{\mathbb{P}_\vartheta}, \mu_{Q_{n,\vartheta}} \right\rangle_{\mathcal{H}} = \left\langle \frac{1}{n}\sum_{i=1}^n k(\cdot, x_i), \mu_{\mathbb{P}_\vartheta} \right\rangle_{\mathcal{H}} =$$

$$= E_{X\sim\mathbb{P}_\vartheta}\left[\frac{1}{n}\sum_{i=1}^n k(X, x_i)\right] = \frac{1}{n}\sum_{i=1}^n E_{X\sim\mathbb{P}_\vartheta}[k_{x_i}(X)]$$

$$\left\langle \mu_{\mathbb{P}_\vartheta}, \mu_{\mathbb{P}_\vartheta} \right\rangle = E_{Y\sim\mathbb{P}_\vartheta}\left[E_{X\sim\mathbb{P}_\vartheta}\left[k(X, Y)\right]\right]$$

First, for $(1) \to 0$ we have:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{n}\sum_{j=1}^{n}k_{x_i}(x_j)-\frac{1}{n}\sum_{i=1}^{n}E_{X\sim\mathbb{P}_\vartheta}[k_{x_i}(X)]=$$
$$=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n}\sum_{j=1}^{n}k_{x_i}(x_j)-E_{X\sim\mathbb{P}_\vartheta}[k_{x_i}(X)]\right)$$

Here, since k_{x_i} are $\mathbb{P}_\vartheta$-measurable, and $E_{X\sim\mathbb{P}_\vartheta}[k_{x_i}(X)]<\infty$, by the strong law of large numbers we have $\left(\frac{1}{n}\sum_{j=1}^{n}k_{x_i}(x_j)-E_{X\sim\mathbb{P}_\vartheta}[k_{x_i}(X)]\right)\to 0$ for all $i\in[n]$, therefore $(1)\to 0$ holds.

Next, $(2)\to 0$ is equivalent to

$$\left\langle\frac{1}{n}\sum_{i=1}^{n}k_{x_i},\mu_{\mathbb{P}_\vartheta}\right\rangle_{\mathcal{H}}-\langle\mu_{\mathbb{P}_\vartheta},\mu_{\mathbb{P}_\vartheta}\rangle_{\mathcal{H}}=\left\langle\frac{1}{n}\sum_{i=1}^{n}k_{x_i}-\mu_{\mathbb{P}_\vartheta},\mu_{\mathbb{P}_\vartheta}\right\rangle_{\mathcal{H}}$$

tending to 0 as $n\to 0$. For this it is more than enough to show that $\langle\frac{1}{n}\sum_{i=1}^{n}k_{x_i}-\mu_{\mathbb{P}_\vartheta},h\rangle_{\mathcal{H}}\to 0\ \forall h\in\mathcal{H}$, i.e. it is the null vector:

$$\left\langle\frac{1}{n}\sum_{i=1}^{n}k_{x_i}-\mu_{\mathbb{P}_\vartheta},h\right\rangle_{\mathcal{H}}=\frac{1}{n}\sum_{i=1}^{n}\langle k_{x_i},h\rangle-\langle\mu_{\mathbb{P}_\vartheta},h\rangle=$$
$$=\frac{1}{n}\sum_{i=1}^{n}h(x_i)-E_{X\sim\mathbb{P}_\vartheta}[h(X)]$$

here, once again, since $E_{X\sim\mathbb{P}_\vartheta}[h(X)]<\infty$ for every element h of the RKHS $\mathcal{H}$, the law of large numbers hold, and this difference tends to 0 as $n\to\infty$. □

Corollary V.2. *From the previous proposition we have* $||\mu_{\mathbb{P}_\vartheta}-\mu_{Q^{(i)}_{n,\vartheta}}||^2_{\mathcal{H}}\to 0$ *for all* $1\le i\le m$ *and* $||\mu_{\mathbb{P}_{\vartheta^*}}-\mu_{Q^{(0)}_{n,\vartheta}}||^2_{\mathcal{H}}\to 0$ *It follows that* $||\mu_{\mathbb{P}_\vartheta}-\mu_{Q^{(0)}_{n,\vartheta}}||^2_{\mathcal{H}}\to||\mu_{\mathbb{P}_\vartheta}-\mu_{\mathbb{P}_{\vartheta^*}}||^2_{\mathcal{H}}>0$, *therefore* $\mathcal{R}(\vartheta)\to 1$ *as* $n\to\infty$ *for all* $\vartheta\neq\vartheta^*$.

Next, to describe the asymptotics in $m\to\infty$ we will use empirical distribution function of the reference variables:

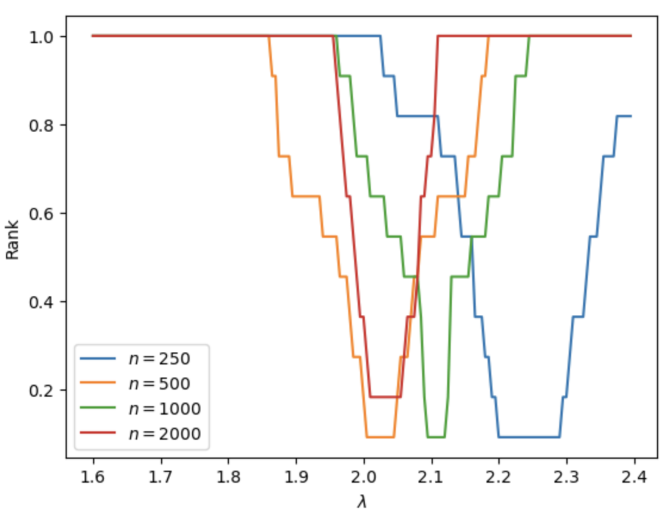

Fig. 2. Rank of the original sample (from an exponential distribution with parameter $\lambda=2$) for $m=10$ resamplings from a given parameter λ using MMD based reference variables.

Definition V.3. Let $X=\{x_1,\dots x_m\}$ be a sample of size m from distribution $\mathbb{P}$. We denote this sample's empirical distribution function with

$$F_{X,m}(x)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}_{x_i<x}$$

Remark. The Ranking function $\mathcal{R}$ can be expressed with the empirical distribution function defined by the reference variables $Z^m(\vartheta)=\{Z^{(i)}(\vartheta)\}_{i=1}^{m-1}$:

$$\mathcal{R}(\vartheta)=\frac{1}{m}\left(1+\sum_{j=1}^{m-1}\mathbb{I}_{\{Z^{(j)}(\vartheta)<Z^{(0)}(\vartheta)\}}\right)=$$
$$=\frac{1}{m}+\frac{m-1}{m}\frac{1}{m-1}\sum_{j=1}^{m-1}\mathbb{I}_{\{Z^{(j)}(\vartheta)<Z^{(0)}(\vartheta)\}}=$$
$$=\frac{1}{m}+\frac{m-1}{m}F_{Z^m(\vartheta),m-1}(Z^{(0)}(\vartheta))$$

Proposition V.4. *From this and the Glivenko-Cantelli Lemma it follows that*

$$\lim_{m\to\infty}\mathcal{R}(\vartheta)=F_{Z(\vartheta)}(Z^{(0)}(\vartheta))$$

where $F_{Z(\vartheta)}$ *is the distribution function of* $Z^{(i)}(\vartheta)$ *(note that these are identically distributed for all* $i\neq 0$*)*

Corollary V.5. *Since* $|\tilde{\mathcal{R}}(\vartheta)-\mathcal{R}(\vartheta)|\le\frac{1}{m}$ *by construction, if* $Z^{(0)}(\vartheta)\le Z_*^{(m-1)}(\vartheta)$ *(i.e.* $\tilde{\mathcal{R}}\le 1$*), the equality above holds for* $\tilde{\mathcal{R}}(\vartheta)$ *as well.*

Remark. For $\bar{\mathcal{R}}$, if $Z^{(i)}(\vartheta)$ are based on one of the MMD constructions, then since $\mathrm{MMD}^2_{\mathcal{H}}[\mathbb{P}_\vartheta,\mathbb{P}_\vartheta]=0$, we have

$$\mathcal{R}(\vartheta)=\lim_{m\to\infty}\bar{\mathcal{R}}(\vartheta)=\lim_{m\to\infty}\left(Z^{(0)}(\vartheta)-Z_*^{(1)}(\vartheta)\right)=$$
$$=Z^{(0)}(\vartheta)-\inf\left\{Z_*^{(1)}(\vartheta)\right\}=Z^{(0)}(\vartheta)$$

VI. Conclusion and Future Work

The framework discussed in this report is a method to tune generative models to find their optimal parameters. Under generative models, we mean the problem where a sample is given from a parameterized distribution, and we want to find the parameters that describe the best this distribution in order to generate synthetic data. It doesn't provide the method to generate the synthetic data itself, for that other (even black box) models, such as neural networks can be used. This framework instead provides a mathematical guarantees for the resulting parameters.

The advantage of this framework lies in it's generality. For example the MMD based reference variables make very little assumptions about the family of distributions that they are working with. The choice of kernels for the MMD based reference variables, or even the reference variables are highly customisable (given that their continuity is ensured).

The contribution of this report to the topic is the proof of continuity for reference variables and the theoretical discussion of the asymptotic behaviors of the rank functions.

Further work to be made in this project include stepwise optimization algorithms to find the estimated parameters and the discussion of their behaviors.

References

[1] Ádám Jung, "Hypothesis test based estimation - technical report," 2024.

[2] A. Tamás and B. C. Csáji, "Distribution-free inference for the regression function of binary classification," *arXiv preprint arXiv:2308.01835*, 2023.

[3] V. I. Paulsen and M. Raghupathi, *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge University Press, 2016, vol. 152.

[4] K. Muandet, K. Fukumizu, B. Sriperumbudur, B. Schölkopf *et al.*, "Kernel mean embedding of distributions: A review and beyond," *Foundations and Trends® in Machine Learning*, 2017.

[5] A. Tamás and B. C. Csáji, "Exact distribution-free hypothesis tests for the regression function of binary classification via conditional kernel mean embeddings," *IEEE Control Systems Letters*, vol. 6, pp. 860–865, 2022.