Ranking Function Based Parameter Estimation

Benedek B. Novák
Supervisor: Balázs Csanád Csáji

I. Introduction

In this report, we elaborate on the previous semerter's work by discussing the asymptotic behaviour of ranking functions, and running some simulations for paramer estimation. For the definitions of *Reproducing Kernel Hilbert Spaces* and *Maximum Mean Discrepancy (MMD)*, see the previous semester's work or [1].

II. The Resampling Framework

We start by giving a brief introduction on the resampling framework. Let $\mathcal{P} = \{\mathbb{P}_\theta \in \Theta\}$ be a class of probability distributions over $\mathcal{X}$ standard Borel space parameterized by Θ polish space. We assume that there is a distribution $\mathbb{P}_{\theta^*} \in \mathcal{P}$, from which we receive a sample $\mathcal{S}^{(0)} \in \mathcal{X}$. ($\mathcal{S}^{(0)}$ can be thought of as a vector of n i.i.d. variables if we have n i.i.d. samples from $\mathbb{P}_{\theta^*}$. However, for most of our purposes, the i.i.d. assumption doesn't need to hold. for example, $\mathcal{S}^{(0)}$ could be a time series as well.)

We also assume, that we have access to a black box G, that can generate a fix sample S given parameter $\theta \in \Theta$ and seed $\xi \in \mathcal{Q}$ where $\mathcal{Q}$ is a standard Borel space, and S has distribution $\mathbb{P}_\theta$ if the seed is drawn from distribution Q over $\mathcal{Q}$. Without loss of generality, it can be assumed that $\mathcal{Q} = [0, 1]$ and Q is the uniform distribution over $\mathcal{Q}$ [2].

Remark. Examples for black box G can be the inverses of the cumulative distribution functions, or neural networks that given random noise can generate meaningful samples. (For example diffusion models for image generation.)

Note that i-th sample can be thought of as a a function of θ if we fix ξ a priori.

The goal is to construct hypothesis tests for $H_0 : \mathbb{P}_\theta = \mathbb{P}_{\theta^*}$ and $H_1 : \mathbb{P}_\theta \neq \mathbb{P}_{\theta^*}$. with exact probability on the type I. error. The main idea of the framework is to generate $m-1$ i.i.d. alternative samples, each from $\mathbb{P}_\theta$ in order to perform the hypothesis test. We denote the original sample with $\mathcal{S}^{(0)}$, and the i-th alternative sample with $S_\theta^{(i)}$. These samples are then compared using a *ranking function*:

Definition II.1. [3] 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 $(a_1, ..., a_m) \in \mathbb{A}^m$:

$$\psi(a_1, a_2, ..., a_m) = \psi(a_1, a_{\pi(2)}, ..., a_{\pi(m)}) \tag{II.1}$$

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

$$\psi(a_i, \{a_k\}_{k\neq i}) \neq \psi(a_j, \{a_k\}_{k\neq j}) \tag{II.2}$$

where the shorthand notation is justified by P1.

Using the concept of ranking functions, we can construct confidence regions for the parameter θ^*:

Theorem II.2. *[4] Given a ranking function ψ, a parameter set Θ, and integer hyperparameters (q, m) with $1 \leq q \leq m$, under the null hypothesis $H_0 : \mathbb{P}_\theta = \mathbb{P}_{\theta^*}$ a confidence region for θ^* can be constructed as:*

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

where we have

$$\mathbb{P}(\theta^* \in \Theta^{\psi}_{(q,m)}) = \frac{q}{m} \tag{II.3}$$

Ranking functions can be defined using a *reference variable* of the original sample:

$$Z_\theta^{(0)} := T(S^{(0)}, \theta) \tag{II.4}$$

where $T : \mathcal{X}^n \times \Theta \to \mathbb{R}$. We can also apply the same function the the alternative samples to obtain $\{Z_\theta^{(i)}\}_{i\neq 0}$ This notion of a reference variables might seem a bit arbitrary at first, so let's have a look at an example:

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

$$Z_\theta^{(i)} = ||\nabla_\theta \mathcal{L}(\theta, S^{(i)})||^2 \tag{II.5}$$

We generalise the concept of reference variables a bit further, introducing a seed component ξ as well, which will be used to create MMD based reference variables. This ξ can be anything from a Borel-measurable space, sampled from an arbitrary distribution, but we will suppose without loss of generality that it is from the $[0, 1]$ interval, and is obtained from a uniform distribution.

$$Z_{\theta,\xi}^{(i)} := T(S_\theta^{(i)}, \theta, \xi_i) \tag{II.6}$$

This generalisation will allow us to use MMD based reference variables defined as:

$$Z_\theta^{(i)} = \widehat{\mathrm{MMD}}^2_{\mathcal{H}}[S_\theta^{(i)}, S_\theta^{(m+i)}] \tag{II.7}$$

Here, we compare all samples to another set of samples $S^{(m)}, S_\theta^{(m+1)}, ..., S_\theta^{(2m-1)}$ to obtain $Z_\theta^{(0)}, ..., Z_\theta^{(m-1)}$, and the seeds encode the random noise that is used the generate the other sample.

Remark. We can think of $\{Z_{\theta,\xi}^{(i)}\}_{i\neq 0}$ as i.i.d. alternative samples for the reference variable $Z_\theta^{(0)}$, which also has the same distribution under H_0. Therefore we will use the notation $Z_\theta^{(i)}$ for this type of reference variable as well, and write out ξ only when we fix the seed.

In order to obtain the rank of the original sample using its reference variable, $Z_\theta^{(0)}, ..., Z_\theta^{(m-1)}$ are sorted in ascending order, therefore the rank of $S_\theta^{(i)}$ becomes its place in the ordering, i.e.

$$\psi(S_\theta^{(i)}, \{S_\theta^{(j)}\}_{j\neq i}) = 1 + \sum_{j\neq i} \mathbb{I}_{\{Z_\theta^{(j)} < Z_\theta^{(i)}\}} \tag{II.8}$$

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

Definition II.3. [3] 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 $Z_\theta^{(i)} <_\pi Z_\theta^{(j)}$ if $Z_\theta^{(i)} < Z_\theta^{(j)}$ or $Z_\theta^{(i)} = Z_\theta^{(j)}$ and $\pi(i) < \pi(j)$.

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

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

$$\mathcal{R}_\theta^{(m)} = 1 + \sum_{i=1}^{m-1} \mathbb{I}_{\{Z_\theta^{(i)} < Z_\theta^{(0)}\}} \tag{II.9}$$

III. Asymptotic behavior

An interesting question that can be asked is what happens if we increase the number of subsamplings (m) or the number of elements in each sample (n). From now on, $\mathcal{R}_\theta^{(m)}$ will denote the *relative rank* of $Z_\theta^{(0)}$, which is the rank divided by m. This will allow us to compare the values of $\mathcal{R}_\theta^{(m)}$ for different ms, as $\mathcal{R}_\theta^{(m)} \in [0,1]$ for every m.

First, we discuss the asymptotics in $m \to \infty$. For this, we define the the relative rank of $z \in \mathbb{R}$ as

$$\mathcal{R}_\theta^{(m)}(z) = \frac{1}{m}\left(1 + \sum_{i=1}^{m-1} \mathbb{I}_{\{Z_\theta^{(i)} < z\}}\right) \tag{III.1}$$

Remark. The Ranking function $\mathcal{R}_\theta^{(m)}(z)$ can be expressed in the following form:

$$\begin{aligned}\mathcal{R}_\theta^{(m)}(z) &= \frac{1}{m}\left(1 + \sum_{i=1}^{m-1} \mathbb{I}_{\{Z_\theta^{(i)} < z\}}\right) = \\ &= \frac{1}{m} + \frac{m-1}{m}\frac{1}{m-1}\sum_{i=1}^{m-1} \mathbb{I}_{\{Z_\theta^{(i)} < z\}}\end{aligned} \tag{III.2}$$

From which, because $\{Z_\theta^{(i)}\}_{i\neq 0}$ are i.i.d., by the application of the law of large numbers, we get

$$\begin{aligned}\lim_{m\to\infty} \mathcal{R}_\theta^{(m)}(z) &= \lim_{m\to\infty} \frac{1}{m-1}\sum_{i=1}^{m-1} \mathbb{I}_{\{Z_\theta^{(i)} < z\}} \\ &= \mathbb{P}\left(Z_\theta^{(1)} < z\right) = F_{Z_\theta^{(1)}}(z)\end{aligned} \tag{III.3}$$

with probability one for a fix $z \in \mathbb{R}$, where $F_{Z_\theta^{(1)}}$ denotes the cummulative distribution function of $Z_\theta^{(1)}$. We can also notice that the relative rank $\mathcal{R}_\theta^{(m)}(z)$ corresponds to the empirical CDF of $\{Z_\theta^{(i)}\}_{i\neq 0}$ at point $z \in \mathbb{R}$, since they are i.i.d..

Corollary III.1. *With the substitution $z = Z_\theta^{(0)}$ we have*

$$\lim_{m\to\infty} \mathcal{R}_\theta^{(m)} = F_{Z_\theta^{(1)}}\left(Z_\theta^{(0)}\right) \tag{III.4}$$

with probability one.

This means that for any parameter θ, the rank of the original sample will converge to the value that the CDF of $Z_\theta^{(1)}$ assigns to the reference variable of the original sample.

Corollary III.2. *Under H_0: $\mathbb{P}_\theta = \mathbb{P}_{\theta^*}$, if $Z_\theta^{(i)}$ are continuous random variables, then $\lim_{m\to\infty} \mathcal{R}_\theta^{(m)} = F_{Z_\theta^{(0)}}\left(Z_\theta^{(0)}\right)$ is uniformly distributed over $[0,1]$.*

For a fix seed ξ, the rank of the reference variable given a parameter θ will be a piecewise constant function (if the distribution is parameterized reasonably). On which it would be difficult to optimize using gradient descent methods, therefore we introduce the *smoothed rank*, which interpolates using the ordered version of $\{Z_\theta^{(i)}\}_{i\neq 0}$ at each point $\theta \in \Theta$. We prove the continuity of the smoothed rank in the next section, now we only show that it's asymptotics behave similarly to that of the relative rank.

Definition III.3. Let (Θ, d) be a metric space, and $Z^{(i)} : \Theta \to \mathbb{R}$ $(i \in [m])$ continuous functions. $Z_*^{(i)}$ is their *pointwise ordered version* if:

$$Z_*^{(i)}(\theta) = \min_{j\in[m]}\left\{Z^{(j)}(\theta) \mid \#\left\{k \mid Z^{(j)}(\theta) \geq Z^{(k)}(\theta)\right\} \geq i\right\} \tag{III.5}$$

i.e. $Z_*^{(1)}(\theta) \leq ... \leq Z_*^{(m)}(\theta)$. (# denotes the cardinality of the set.)

Definition III.4. Let $Y_\theta^{(1)} \leq ... \leq Y_\theta^{(m-1)}$ denote the pointwise ordered version of $\{Z_\theta^{(i)}\}_{i\neq 0}$. Then the *smoothed rank* of $z \in \mathbb{R}$ is defined as:

$$\tilde{\mathcal{R}}_{\theta,\xi}^{(m)}(z) = \begin{cases} \frac{1}{m}\left(\frac{z}{Y_\theta^{(1)}}\right) & \text{if } z < Y_\theta^{(1)} \\ \frac{1}{m}\left(k + \frac{z - Y_\theta^{(k)}}{Y_\theta^{(k+1)} - Y_\theta^{(k)}}\right) & \text{if } Y_\theta^{(k)} \leq z < Y_\theta^{(k+1)} \\ \frac{1}{m}\left(m + \tau\left(z, Y_\theta^{(m-1)}\right)\right) & \text{if } Y_\theta^{(m-1)} \leq z \end{cases} \tag{III.6}$$

where τ is a continuous function with $\tau(z,y) \geq 0$ and $\tau(z,z) = 0$ for every z and y in the ranges of $Z_\theta^{(0)}$ and $Z_\theta^{(1)}$, assuming $z \geq y$. Furthermore, we require τ to monotonically increase in z and monotonically decrease in y in the same area.

The selection of τ can be used to adjust the slope in this region of the function during optimization, in order to find the region where $\mathcal{R}_\theta^{(m)} < 1$. Examples of the choice of τ can be $\tau(z,y) = \frac{z}{y} - 1$ or $\tau(z,y) = \frac{z^2}{y^2} - 1$

Similarly to $\mathcal{R}_\theta^{(m)}$, we define $\tilde{\mathcal{R}}_\theta^{(m)} = \tilde{\mathcal{R}}_\theta^{(m)}(Z_\theta^{(0)})$.

Theorem III.5. *Let $\tilde{\mathcal{R}}_\theta^{(m)}(z)$ denote the smoothed rank of $z \in \mathbb{R}$. Then*

$$\lim_{m\to\infty} \tilde{\mathcal{R}}_\theta^{(m)}(z) = \mathbb{P}\left(Z_\theta^{(1)} < z\right) = F_{Z_\theta^{(1)}}(z) \tag{III.7}$$

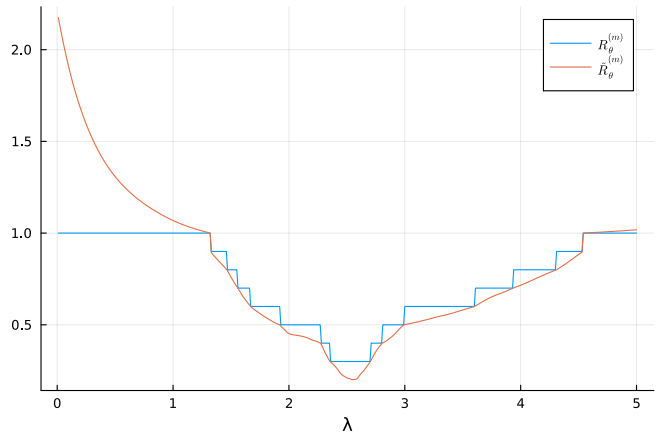


Fig. 1: The relative rank and smoothed rank (for a fix seed) of a sample from an exponential distribution with parameter 2 using MMD based reference variables using the RBF kernel with $\sigma = 1$. ($n = 50$, $m = 10$, $\tau(z, y) = \frac{z}{y} - 1$)

Proof. Let $Z_\theta^{(\max)}$ denote the maximum of $\{Z_\theta^{(i)}\}_{i\neq 0}$ and

$$c = \mathbb{P}\left(Z_\theta^{(1)} \geq z\right) \tag{III.8}$$

First, we assume $c > 0$. Notice that for every $z \in \mathbb{R}$, if $z \leq Z_\theta^{(\max)}$, then by construction $|\tilde{\mathcal{R}}_\theta^{(m)}(z) - \mathcal{R}_\theta^{(m)}(z)| \leq \frac{1}{m}$. This means that if $z \leq Z_\theta^{(\max)}$ for a large enough m, then the limit of $\tilde{\mathcal{R}}_\theta^{(m)}(z)$ and $\mathcal{R}_\theta^{(m)}(z)$ will be the same. Since we already know that $\mathcal{R}_\theta^{(m)}(z)$ converges pointwise a.s. to the cdf (III.3), the probability of $z \leq Z_\theta^{(\max)}$ for a large enough m will give a lower bound for the probability of the convergence in (III.7):

$$\begin{aligned}
\mathbb{P}\Big(\lim_{m\to\infty} \tilde{\mathcal{R}}_\theta^{(m)}(z) &= F_{Z_\theta^{(1)}}(z)\Big) \\
&\geq \mathbb{P}\left(z \leq \lim_{m\to\infty} Z_\theta^{(\max)}\right) \\
&= 1 - \mathbb{P}\left(z > \limsup_{i\to\infty} Z_\theta^{(i)}\right) \\
&= 1 - \prod_{i=1}^{\infty} \mathbb{P}\left(z > Z_\theta^{(i)}\right) \\
&= 1 - \prod_{i=1}^{\infty} (1 - c) = 1
\end{aligned}$$

since $\mathbb{P}\left(z > Z_\theta^{(i)}\right) = 1 - c < 1$.

If $c = 0$, then it means that $Z_\theta^{(\max)} < z$ almost surely. From the definition of the relative smoothed rank, we can see that

$$\tilde{\mathcal{R}}_\theta^{(m)}(z) = \frac{1}{m}\left(m + \tau\left(z, Z_\theta^{(\max)}\right)\right) \tag{III.9}$$

almost surely. Since we required τ to be monotonically decreasing in its second argument, $f(z, Z_\theta^{(\max)}) \leq \tau(z, Z_\theta^{(1)})$.

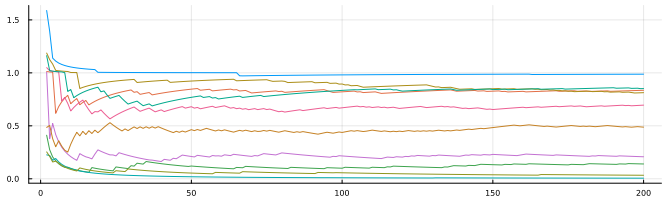


Fig. 2: The smoothed rank in θ^* for 10 different fixed seeds with increasing m.

This can be used to give an upper bound to the limit:

$$\begin{aligned}
\lim_{m\to\infty} \tilde{\mathcal{R}}_\theta^{(m)}(z) &= \lim_{m\to\infty} \frac{1}{m}\left(m + \tau\left(z, Z_\theta^{(\max)}\right)\right) \\
&= 1 + \lim_{m\to\infty} \frac{\tau\left(z, Z_\theta^{(\max)}\right)}{m} \\
&\leq 1 + \lim_{m\to\infty} \frac{\tau\left(z, Z_\theta^{(1)}\right)}{m} = 1
\end{aligned} \tag{III.10}$$

Since we required $\tau(z, y) \geq 0$, $\tilde{\mathcal{R}}_\theta^{(m)}(z) \geq 1$ also holds, therefore $\lim_{m\to\infty} \tilde{\mathcal{R}}_\theta^{(m)}(z) = 1$, which is exactly the value that $F_{Z_\theta^{(1)}}(z)$ would take if $Z_\theta^{(1)} < z$ almost surely.

□

Corollary III.6. *With the substitution $z = Z_\theta^{(0)}$ we have*

$$\lim_{m\to\infty} \tilde{\mathcal{R}}_\theta^{(m)} = F_{Z_\theta^{(1)}}\left(Z_\theta^{(0)}\right) \tag{III.11}$$

with probability one for any fixed seed ξ.

Remark. Under the null hypothesis $\mathbb{P}_\theta = \mathbb{P}_{\theta^*}$, if $Z_\theta^{(i)}$ are continuous random variables, then $\lim_{m\to\infty} \tilde{\mathcal{R}}_\theta^{(m)} = F_{Z_\theta^{(0)}}\left(Z_\theta^{(0)}\right)$ is uniformly distributed over $[0, 1]$.

Next, we investigate the asymptotic behavior for $n \to \infty$. For this, we assume that the original sample $S^{(0)} = \{x_1, ..., x_n\}$ contains i.i.d. instances from distribution $\mathbb{P}_\theta$.

Definition III.7. We say that a reference variable is *consistent*, if it holds that

$$\lim_{n\to\infty} Z_\theta^{(i)} = \begin{cases} 0 & \text{if } x_j \sim \mathbb{P}_\theta \text{ i.i.d.} \\ c \in \mathbb{R}_+ \cup \{\infty\} & \text{else} \end{cases} \tag{III.12}$$

almost surely for any $\theta \in \Theta$ parameter.

Proposition III.8. *(Pointwise consistency) If $Z_\theta^{(i)}$ are consistent, then for any fix $\theta \in \Theta$, the relative rank $\mathcal{R}_\theta^{(m)}$ constructed from it has the following properties: I.) $\tilde{\mathcal{R}}_\theta^{(m)} \to 1$ a.s. as $n \to \infty$ if $\mathbb{P}_{\theta^*} \neq \mathbb{P}_\theta$.*

II.) $\tilde{\mathcal{R}}_\theta^{(m)} \xrightarrow{d} U_m[0, 1]$ as $n \to \infty$ if $\mathbb{P}_{\theta^*} = \mathbb{P}_\theta$ where $U_m[0, 1]$ denotes the discrete uniform distribution over the set $\left\{\frac{1}{m}, ..., \frac{m-1}{m}, 1\right\}$.

Proof. I.) Since $\mathbb{P}_\theta \neq \mathbb{P}_{\theta^*}$, by the definition of consistency, we have $\lim_{n\to\infty} Z_\theta^{(0)} = c > 0 = \lim_{n\to\infty} Z_\theta^{(i)}$ for all $i \neq 0$, therefore $\lim_{n\to\infty} \tilde{\mathcal{R}}_{\theta,\xi}^{(m)} = 1$ a.s..

II.) Since $\mathbb{P}_{\theta^*} = \mathbb{P}_\theta$, $Z_\theta^{(0)}$ and $Z_\theta^{(i)}$ have the same distribution for all $i \in [m]$, the place that the reference variable would take in the ordering of $\{Z_\theta^{(i)}\}$ is uniformly distributed. □

Next, we prove that the MMD based reference variables are consistent if they are constructed using a characteristic kernel. For this, we need the law of large numbers for kernel mean embeddings:

Lemma III.9. *(Law of large numbers for kernel mean embeddings) Let $\mathcal{H}$ be a real RKHS over $\mathcal{X}$ and $(\mathcal{X}, \mathcal{A}, \mathbb{P})$ a probability space. Denote the empirical distribution of a sample $\{x_i\}_{i=1}^n$ with $Q_n(A) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{x_i \in A}$ for all $A \in \mathcal{A}$. Then it holds that $||\mu_{\mathbb{P}} - \mu_{Q_n}||_{\mathcal{H}}^2 \to 0$ as $n \to \infty$.*

Proof. First we decompose the norm into two parts, and then show that each converges to 0 as $n \to 0$:

$$\begin{aligned} ||\mu_{\mathbb{P}} - \mu_{Q_n}||_{\mathcal{H}}^2 &= \langle \mu_{\mathbb{P}} - \mu_{Q_n}, \mu_{\mathbb{P}} - \mu_{Q_n} \rangle_{\mathcal{H}} \\ &= \langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle - 2\langle \mu_{\mathbb{P}}, \mu_{Q_n} \rangle + \langle \mu_{Q_n}, \mu_{Q_n} \rangle \\ &= (\langle \mu_{Q_n}, \mu_{Q_n} \rangle - \langle \mu_{\mathbb{P}}, \mu_{Q_n} \rangle) \qquad \text{(III.13)} \\ &\quad + (\langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle - \langle \mu_{\mathbb{P}}, \mu_{Q_n} \rangle) \qquad \text{(III.14)} \end{aligned}$$

We can rewrite the scalar products in (III.13) and (III.14) using the definition of the kernel mean embeddings and the reproducing property of the RKHS:

$$\begin{aligned} \langle \mu_{Q_n}, \mu_{Q_n} \rangle_{\mathcal{H}} &= \left\langle \frac{1}{n}\sum_{i=1}^n k(\cdot, x_i), \frac{1}{n}\sum_{j=1}^n k(\cdot, x_j) \right\rangle_{\mathcal{H}} \\ &= \frac{1}{n}\sum_{i=1}^n \frac{1}{n}\sum_{j=1}^n k_{x_i}(x_j) \\ \langle \mu_{\mathbb{P}}, \mu_{Q_n} \rangle_{\mathcal{H}} &= \left\langle \frac{1}{n}\sum_{i=1}^n k(\cdot, x_i), \mu_{\mathbb{P}} \right\rangle_{\mathcal{H}} \\ &= E_{X\sim\mathbb{P}}\left[\frac{1}{n}\sum_{i=1}^n k(X, x_i)\right] \\ &= \frac{1}{n}\sum_{i=1}^n E_{X\sim\mathbb{P}}[k_{x_i}(X)] \\ \langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle &= E_{Y\sim\mathbb{P}}\left[E_{X\sim\mathbb{P}}\left[k(X, Y)\right]\right] \end{aligned}$$

First, for (III.13) $\to 0$ we have:

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

Next we use the fact that k_{x_i} are $\mathbb{P}$-measurable functions with real values and $E_{X\sim\mathbb{P}}[k_{x_i}(X)] = \langle k_{x_i}, \mu_{\mathbb{P}} \rangle \leq ||k_{x_i}||_{\mathcal{H}} ||\mu_{\mathbb{P}}||_{\mathcal{H}} < \infty$, therefore the strong law of large numbers can be applied for each $i \in [n]$:

$$\left(\frac{1}{n}\sum_{j=1}^n k_{x_i}(x_j) - E_{X\sim\mathbb{P}}[k_{x_i}(X)]\right) \to 0$$

so their sum (III.13) $\to 0$ as well.

Next, (III.14) $\to 0$ is equivalent to

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

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

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

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

Definition III.10. A kernel function k is a *characteristic kernel*, if the corresponding kernel mean embedding captures all information about the underlying distributions:

$$||\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}||_{\mathcal{H}}^2 = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q} \qquad \text{(III.15)}$$

i.e. the MMD of two embedded distributions is 0 if and only if they are the same distribution.

Corollary III.11. *If an MMD-based reference variable is constructed using a characteristic kernel, then it is consistent.*

Proof. Let $X_n = \{x_1, ..., x_n\}$ be the original sample, and $X_n' = \{x_1', ..., x_n'\}$ be the sample that is drawn using seed ξ for the calculation of the reference variable. We denote their empirical distributions as Q_n and Q_n' respectively. From the previous lemma, we have that $||\mu_{\mathbb{P}_{\theta^*}} - \mu_{Q_n}||_{\mathcal{H}}^2 \to 0$ and $||\mu_{\mathbb{P}_\theta} - \mu_{Q_n'}||_{\mathcal{H}}^2 \to 0$ as $n \to \infty$. Therefore

$$\begin{aligned} \widehat{\text{MMD}}_{\mathcal{H}}^2[X_n, X_n'] &= ||\mu_{Q_n} - \mu_{Q_n'}||_{\mathcal{H}}^2 \\ &= ||\mu_{Q_n} - \mu_{\mathbb{P}_{\theta^*}} + \mu_{\mathbb{P}_{\theta^*}} - \mu_{\mathbb{P}_\theta} + \mu_{\mathbb{P}_\theta} - \mu_{Q_n'}||_{\mathcal{H}}^2 \\ &\leq ||\mu_{Q_n} - \mu_{\mathbb{P}_{\theta^*}}||_{\mathcal{H}}^2 + ||\mu_{\mathbb{P}_{\theta^*}} - \mu_{\mathbb{P}_\theta}||_{\mathcal{H}}^2 + ||\mu_{\mathbb{P}_\theta} - \mu_{Q_n'}||_{\mathcal{H}}^2 \end{aligned}$$

We can see that an upper bound for the limit is

$$\lim_{n\to\infty} \widehat{\text{MMD}}_{\mathcal{H}}^2[X_n, X_n'] \leq \text{MMD}_{\mathcal{H}}^2[\mathbb{P}_{\theta^*}, \mathbb{P}_\theta] \qquad \text{(III.16)}$$

By changing the role of $\widehat{\text{MMD}}_{\mathcal{H}}^2$ and $\text{MMD}_{\mathcal{H}}^2$, we get a lower bound as well:

$$\begin{aligned} \text{MMD}_{\mathcal{H}}^2[\mathbb{P}_{\theta^*}, \mathbb{P}_\theta] &= ||\mu_{\mathbb{P}_{\theta^*}} - \mu_{\mathbb{P}_\theta}||_{\mathcal{H}}^2 \\ &= ||\mu_{\mathbb{P}_{\theta^*}} - \mu_{Q_n} + \mu_{Q_n} - \mu_{Q_n'} + \mu_{Q_n'} - \mu_{\mathbb{P}_\theta}||_{\mathcal{H}}^2 \\ &\leq ||\mu_{\mathbb{P}_{\theta^*}} - \mu_{Q_n}||_{\mathcal{H}}^2 + ||\mu_{Q_n} - \mu_{Q_n'}||_{\mathcal{H}}^2 + ||\mu_{Q_n'} - \mu_{\mathbb{P}_\theta}||_{\mathcal{H}}^2 \end{aligned}$$

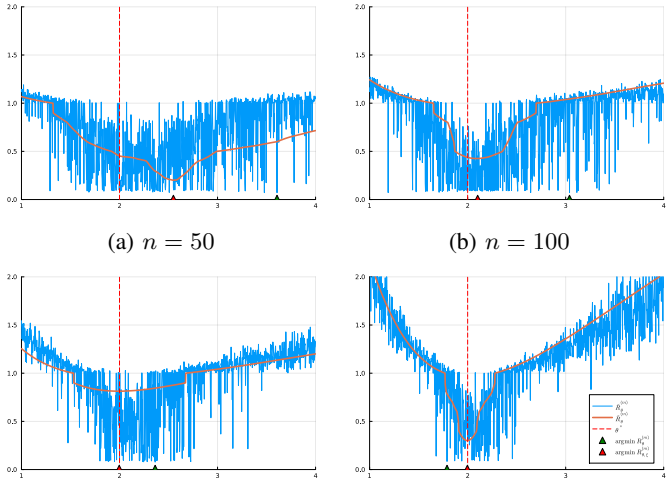
Once again, two of the three terms tend to 0 as $n \to \infty$, therefore

$$\text{MMD}_{\mathcal{H}}^2[\mathbb{P}_{\theta^*}, \mathbb{P}_\theta] \leq \lim_{n\to\infty} \widehat{\text{MMD}}_{\mathcal{H}}^2[X_n, X_n'] \qquad \text{(III.17)}$$

From which we get that

$$\begin{aligned} \lim_{n\to\infty} \widehat{\text{MMD}}_{\mathcal{H}}^2[X_n, X_n'] &= \text{MMD}_{\mathcal{H}}^2[\mathbb{P}_{\theta^*}, \mathbb{P}_\theta] \\ &= 0 \ \text{ if and only if } \mathbb{P}_\theta = \mathbb{P}_{\theta^*} \end{aligned}$$

□



(a) $n = 50$ (b) $n = 100$

(c) $n = 200$ (d) $n = 500$

Fig. 3: The smoothed rank at each parameter with and without fixing the seed. The original sample of size n with i.i.d. instances came from an exponential distribution with paramer $\theta = 2$. ($m = 10$, using RBF kernel MMD based reference variables.)

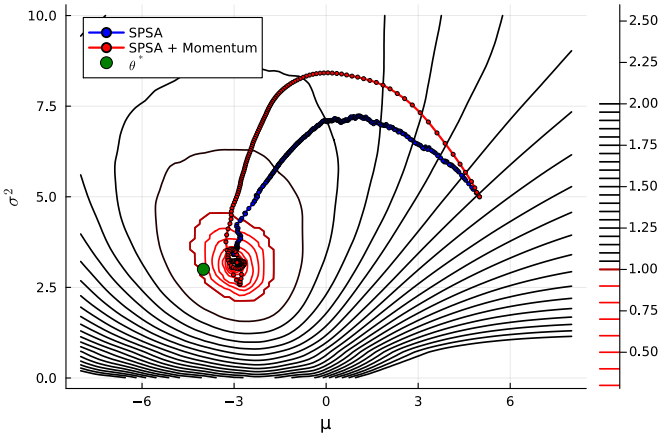


Fig. 4: Kiefer-Wolfowitz based optimization for a sample from $\mathcal{N}(-4, 3), n = 50, m = 20, \tau(z, y) = \frac{z}{y} - 1$ using RBF kernel based MMD reference variables.

IV. Optimization

When looking for a minimizer $\hat{\theta}$ of the smoothed rank, we need to fix the seed from which the alternative samples are drawn from, in order to make $\tilde{\mathcal{R}}^{(m)}_{\theta,\xi}$ a smooth function without any randomness. We use stochastic approximation methods, such as the Kiefer–Wolfowitz algorithm for 1 dimensional parameter spaces, and Simultaneous Perturbation Stochastic Approximation (SPSA) [5] for higher dimensional spaces to find the minimum of $\tilde{\mathcal{R}}^{(m)}_{\theta,\xi}$ in Θ. These algorithm are very similar to the stochastic gradient descent, but the difference is that they estimate the gradient locally by taking a step in each direction, instead of calculating it exactly.

Definition IV.1. Kiefer-Wolfowitz Algorithm for finding a minimum of function $F : \mathbb{R} \to \mathbb{R}$:

$$\theta_{n+1} = \theta_n + \gamma_n \frac{F(\theta_n - \delta_n) - F(\theta_n + \delta_n)}{2\delta_n} \tag{IV.1}$$

where the learning rates $\{\gamma_n\}$ and δ_n satisfy $\sum_{n=0}^{\infty} \gamma_n = \infty$, $\lim_{n\to\infty} \delta_n = 0$ and $\sum_{n=0}^{\infty} \frac{\gamma_n^2}{\delta_n^2} < \infty$.

Definition IV.2. Simultaneous Perturbation Stochastic Approximation (SPSA): for finding a minimum of $F : \mathbb{R}^d \to \mathbb{R}$:

$$\theta_{n+1,k} = \theta_{n,k} + \gamma_n \frac{F(\theta_n - \delta_n \Delta_n) - F(\theta_n + \delta_n \Delta_n)}{2\delta_n \Delta_{n,k}} \tag{IV.2}$$

where $\theta_{n,k}$ denotes the kth coordinate of θ_n, and $\{\Delta_n\}$ are independent, symmetric, zero-mean vectors, for example Bernoulli trials with $\Delta_{n,k} = \pm 1$ with probability $\frac{1}{2}$ each.

V. Conclusion and future work

In this semester I examined the asymptotic behaviour of reference variables, and have given criteria for their pointwise convergence. I have also made some simulations for low dimensional parameter spaces and simple distributions. Further experiments on more compex distributions could be made, for which improvements on the currently used optimization algorithms will be required.

References

[1] 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.

[2] O. Kallenberg and O. Kallenberg, *Foundations of modern probability*. Springer, 1997, vol. 2.

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

[4] 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.

[5] J. Spall, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Transactions on Automatic Control*, vol. 37, no. 3, pp. 332–341, 1992.