Resampling Based Estimation of Generative Models

Benedek Bálint Novák
Supervisor: Balázs Csanád Csáji

I. Introduction

An important problem in machine learning and related fields is the estimation of a probability distribution given an i.i.d. sample from this distribution. For parameterised distribution families there are many well studied methods for creating an estimate for the original parameter, in both classical statistics [1] and in machine learning [2], or in recent years, with the increase in computational capacity and interest in deep neural networks, large language models and diffusion models (for text and image generation) have gained a lot of popularity [3][4]. There are also non-parametric estimation methods, such as kernel density estimation [5]. However, these methods only provide point estimates for the distribution, which differs based on the chosen model. This phenomena (called the Rashomon effect [6]) has led to the argument recently that it might not be sufficient to study single models and their behaviours for high-stakes decisions, for example in medicine or finance [7][8]. Therefore one of our goals will be to establish a framework for having a set of equivalently suitable estimations for distributions.

Another problem that might arise when using classical models is that they require an explicitly computable function of the parameter to see how well it fits the sample. This can be problematic if there is no oversight over the parametrisation and sample generation, or if the likelihood function cannot be calculated explicitly. Therefore, our second goal will be to make these estimations distribution-free. This means that the estimator does not depend on some previous knowledge about the underlying distribution, or even the process that generates the sample based on a given parameter. This approach can useful when we have a black box model that we would like to fine-tune.

In order to achieve these goals, the resampling and ranking based framework [9] will be used, which has already been successfully utilized to create confidence regions with an exact (even for finite sample sizes), user-chosen coverage (probability of the type I error) for the parameters of regression functions of binary classification problems [10]. Other resampling-based methods have been successfully used for non-asymptotic confidence regions for kernelized regression models [11].

II. Kernel Mean Embeddings

A. Reproducing Kernel Hilbert Spaces

For the sake of completeness, we begin our discussion with a brief introduction on Reproducing Kernel Hilbert Spaces and Kernel Mean Embeddings.

Definition II.1. *[12, Definition 1.1] Let $\mathcal{X}$ be an arbitrary set, $\mathcal{H}$ 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, i.e. there exists a $C_x > 0$ for all E_x such that $|E_x(f)| \leq C_x ||f||_{\mathcal{H}}$.*

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

Definition II.2. *[12, Definition 1.2] The reproducing kernel of RKHS $\mathcal{H}$ over $\mathcal{X}$ is the function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined as $k(x, y) := \langle k_y, k_x \rangle_{\mathcal{H}}$.*

This means that for every $x \in \mathcal{X}$, there is a $k_x \in \mathcal{H}$ that represents it in the *feature space* $\mathcal{H}$ and we can compare them using $k(\cdot, \cdot)$ without explicitly having to calculate this map.

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

Now we can see that how having an RKHS $\mathcal{H}$ leads to being able to perform operations on the embedded elements of $\mathcal{X}$ using the *kernel function* k. However, the question of how to find a suitable RKHS $\mathcal{H}$ and the corresponding kernel k still remains. Luckily, the Moore-Aronszajn Theorem shows that if the *kernel function* k is *positive definite*, then there uniquely exists a corresponding RKHS.

Definition II.3. *[12, Chapter 2.2] We say that a symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 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*

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

Theorem II.4 (Moore-Aronszajn)**.** *[12, Theorem 2.14]*
If a symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite, then there uniquely exists a corresponding RKHS $\mathcal{H}$.

B. Kernel Mean Embedding and MMD

Now that Reproducing Kernel Hilbert Spaces are defined, let's have a look at how to embed not only single elements of $\mathcal{X}$ into them, but whole probability distributions over $\mathcal{X}$.

Definition II.5. *[13, Definition 3.1] The kernel mean embedding of a probability measure $\mathbb{P}$ over $\mathcal{X}$ into an RKHS $\mathcal{H}$ with reproducing kernel k is defined as*

$$\mu_{\mathbb{P}} = \int k(\cdot, x) \, \mathrm{d}\mathbb{P}(x) \tag{II.2}$$

Here the integral is to be interpreted as a Bochner-integral, as defined in [13, Chapter 3.1] in a similar manner to the Lebesgue integral.

The following lemma gives us a sufficient condition for the mean embedding $\mu_{\mathbb{P}}$ to be an element of the RKHS $\mathcal{H}$.

Lemma II.6. *[13, Lemma 3.1] If $\mathbb{E}_{X\sim\mathbb{P}}[\sqrt{k(X,X)}] < \infty$, then $\mu_{\mathbb{P}} \in \mathcal{H}$ and $\mathbb{E}_{X\sim\mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}}\rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$.*

This means that we can compare distributions to each other as if they were elements of a Hilbert space, and so the *Maximum Mean Discrepancy* and its unbiased estimator can be defined as follows:

Definition II.7. *[13, equation 3.29] 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:*

$$\begin{aligned}\mathrm{MMD}_{\mathcal{H}}[\mathbb{P},\mathbb{Q}] &= \sup_{||f||\leq 1}\left\{\int f(x)\,d\mathbb{P}(x) - \int f(y)\,d\mathbb{Q}(y)\right\}\\ &= \sup_{||f||\leq 1}\{\langle f, \mu_{\mathbb{P}} - \mu_q\rangle\}_{\mathcal{H}}\\ &= ||\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}||_{\mathcal{H}}\end{aligned} \tag{II.3}$$

Using $X, X' \sim \mathbb{P}$ and $Y, Y' \sim \mathbb{Q}$ independent random variables and the reproducing property, we get:

$$\mathrm{MMD}^2_{\mathcal{H}}[\mathbb{P},\mathbb{Q}] = \mathbb{E}[k(X,X')] + \mathbb{E}[k(Y,Y')] - 2\mathbb{E}[k(X,Y)] \tag{II.4}$$

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 [13, equation 3.32]:

$$\begin{aligned}\widehat{\mathrm{MMD}}^2_{\mathcal{H}}[X,Y] =& \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}k(x_i,x_j)\\ &+ \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j=1}^{m}k(y_i,y_j)\\ &- \frac{2}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n}k(x_i,y_j)\end{aligned} \tag{II.5}$$

There is another useful definition to be mentioned here. If an RKHS is rich enough to represent all probability distributions uniquely, we call its reproducing kernel *characteristic*:

Definition II.8. *[13, Definition 3.2] A kernel function k is a characteristic kernel, if the corresponding kernel mean embedding captures all information about the underlying distributions:*

$$||\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}||^2_{\mathcal{H}} = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q} \tag{II.6}$$

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

Remark. It has been shown that the Gaussian and Laplacian kernels are characteristic on $\mathbb{R}^d$. [14, Theorem 2]

III. The Estimation Framework

A. Resampling and Ranking

The Resampling and Ranking Framework is similar to the one used in classical statistics, in the sense that a statistical field is examined. Let $\mathcal{X}$ be a standard Borel space equiped with its Borel σ-algebra $\mathcal{A}$, and let $\mathcal{P} = \{\mathbb{P}_\theta \in \Theta\}$ be a class of probability distributions over $\mathcal{X}$ (i.e. $(\mathcal{X}, \mathcal{A}, \mathbb{P}_\theta)$ is a probability space for every $\theta \in \Theta$). We assume that there is a distribution $\mathbb{P}_{\theta^*} \in \mathcal{P}$, from which we receive a sample $\mathcal{S}^{(0)} \in \mathcal{X}$.

A slight distinction between the two frameworks is that we assume that only one sample is available from the distribution, instead of the usual i.i.d. sample of size n assumption. Note that the latter is a special case of the former, since in our case $\mathcal{S}^{(0)}$ can be thought of as the vector of the 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 most of the methods discussed could be applied to time series as well)

The other, and most important assumption is that we have access to a black box G, that can generate an *alternative sample* for any given parameter, and its seed can be fixed. More formally, there exists a standard Borel space $\mathcal{Q}$ with probability distribution Q and function $G : \Theta \times \mathcal{Q} \to \mathcal{X}$ that is Borel-measurable for every $\theta \in \Theta$ such that $G(\xi, \theta) \sim \mathbb{P}_\theta$ for every $\theta \in \Theta$.

Without loss of generality, it can be assumed that $\mathcal{Q} = [0,1]$ and Q is the uniform distribution over $\mathcal{Q}$. [15, Theorem A1.6]

Examples for black boxes G can be the inverses of the cumulative distribution functions [16], or neural networks that can generate meaningful samples given random noise, such as diffusion models for image generation [3][4].

The goal is of course to estimate the distribution $\mathbb{P}_{\theta^*}$ that the sample could have come from – not only as a point estimate, but as a whole region of equally suitable parameters. This estimation will be done using hypothesis tests for $H_0 : \mathbb{P}_\theta = \mathbb{P}_{\theta^*}$ and $H_1 : \mathbb{P}_\theta \neq \mathbb{P}_{\theta^*}$ that have exact, user-chosen coverage (probability of type I error), and are strongly consistent. These tests are constructed using the resampling and ranking framework introduced in [9].

The core idea of the framework is to generate $m-1$ i.i.d. alternative samples, each from $\mathbb{P}_\theta$, order them using a *ranking function*, and then the *rank* of the original sample becomes its place in the ordering:

Definition III.1. *[10, Definition 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 $(a_1, ..., a_m) \in \mathbb{A}^m$:

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

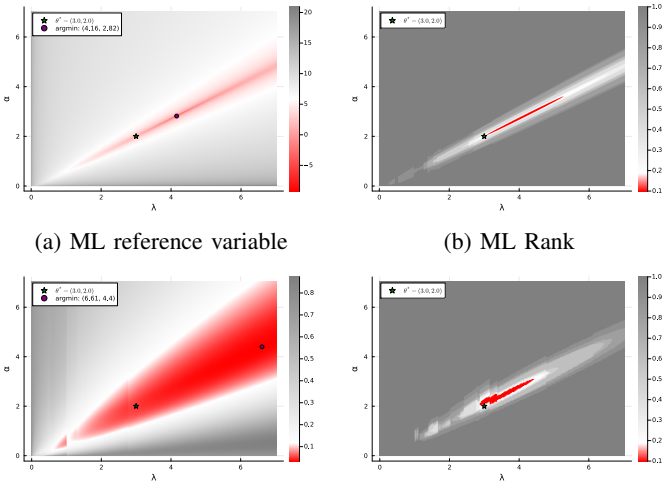
where π is a permutation on the set $\{2, ..., m\}$.
P2) Uniqueness in the first variable, i.e. for all $i, j \in [m]$ if $a_i \neq a_j$, then

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

where the shorthand notation is justified by P1.

Using the concept of ranking functions, confidence regions for parameter θ^* with exact coverage can be constructed. From now on the original sample is denoted with $S^{(0)}$, and the i-th alternative sample with $S^{(i)}_\theta$.

Theorem III.2. *[10, Theorem 2] Given a ranking function ψ, a parameter set Θ and integer hyperparameters (p, q, m) with*



(a) ML reference variable (b) ML Rank

(c) MMD reference variable (d) MMD Rank

Fig. 1: Examples for Loglikelihood- and MMD-based reference variables and their ranks for a sample of size 50 from a Gamma distribution, using a fixed seed to generate both the reference variables and the alternative samples.

$1 \leq p \leq q \leq m$, *for the null hypothesis* $H_0 : \mathbb{P}_\theta = \mathbb{P}_{\theta^*}$ *a confidence region for* θ^* *can be constructed as:*

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

where we have

$$\mathbb{P}\left(\theta^* \in \Theta^{\psi}_{(p,q,m)}\right) = \frac{q-p+1}{m} \tag{III.3}$$

The Rank in the previous example was calculated using *reference variables*, which are functions that assign a value to the original sample based on the parameter it is being tested for. In this work, we will focus on reference variable based ranking functions, for which the reference varaiables will be denoted as:

$$Z^{(0)}_\theta := T(S^{(0)}, \theta) \tag{III.4}$$

where $T : \mathcal{X} \times \Theta \to \mathbb{R}$. The same function can also be applied to the alternative samples in order to obtain $\{Z^{(i)}_\theta\}_{i\neq 0}$. The rank then becomes the position of the reference variable in the ordering of all references.

An example for a reference variable that was used in Figure 1b is the Loglikelihood-based reference variable:

Example. If $\mathcal{L}(\theta, S^{(i)})$ denotes the log-likelihood of sample $S^{(i)}$, then

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

is the *Loglikelihood-based reference variable*.

This type of reference variable is useful if we know that the sample came from a well-known distribution family, for which the expression above can be calculated. Its main advantage is that it is really fast to calculate if we have an explicit solution for the expression above.

However, we would like to focus on *distribution-free* reference variables that don't depend on such prior knowledge or calculations. To achieve this end, the concept of reference variables can be further generalised by introducing some randomization to them through a *seed* component:

$$Z^{(i)}_{\theta,\xi_i} := T(S^{(i)}_\theta, \theta, \xi_i) \tag{III.6}$$

Where the seeds $\xi = \{\xi_i\}_{i=0}^{m-1}$ are sampled i.i.d. from a Borel-measurable space over an arbitrary distribution. Note that if the seed ξ is fixed, then these reference variables act as if they were of the previous type.

This generalisation will allow us to introduce MMD based reference variables using the unbiased estimator for the MMD of the distributions of the two samples:

$$Z^{(i)}_{\theta,\xi_i} = \widehat{\mathrm{MMD}}^2_{\mathcal{H}}[S^{(i)}_\theta, S^{(m+i)}_\theta] \tag{III.7}$$

Here all samples are compared to another set of alternative samples to obtain $Z^{(0)}_\theta, ..., Z^{(m-1)}_\theta$ and ξ_i encodes the seed that is used the generate $S^{(m)}_\theta, S^{(m+1)}_\theta, ..., S^{(2m-1)}_\theta$ using black box $G : \Theta \times \mathcal{Q} \to \mathcal{X}$.

Remark. $\{Z^{(i)}_{\theta,\xi_i}\}_{i\neq 0}$ can be thought of as i.i.d. alternative samples for the reference variable $Z^{(0)}_\theta$, which also have the same distribution under H_0. Therefore the notation $Z^{(i)}_\theta$ is used for this type of reference variable as well, indicating ξ only when the seed is fixed.

In order to obtain the rank of the original sample using its reference variable, $Z^{(0)}_\theta, ..., Z^{(m-1)}_\theta$ are sorted in ascending order, and the rank of $S^{(i)}_\theta$ becomes its place in the ordering:

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

Remark. It can be seen from equation III.8 why the i.i.d. assumption only needs to hold for the resamplings and why it can be relaxed for the original sample. It's because the rank of $S^{(0)}$ and therefore the hypothesis test only depends on the sample through the reference variable, which can be tailored to the task at hand.

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

Definition III.3. *[10, IV/A] 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* $Z^{(i)}_\theta <_\pi Z^{(j)}_\theta$ *if* $Z^{(i)}_\theta < Z^{(j)}_\theta$ *or* $Z^{(i)}_\theta = Z^{(j)}_\theta$ *and* $\pi(i) < \pi(j)$.

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

In order to compare the ranks of the original sample for different numbers of resamplings, we will be using the notion of the *normalized rank* instead. The *normalized rank* of the original sample with regards to the $m-1$ i.i.d samples generated from $\mathbb{P}_\theta$ is the rank devided by the number of resamplings, i.e.:

$$\mathcal{R}^{(m)}_\theta = \frac{1}{m}\left(1 + \sum_{i=1}^{m-1} \mathbb{I}_{\{Z^{(i)}_\theta < Z^{(0)}_\theta\}}\right) \tag{III.9}$$

B. Smoothed Rank

As we saw in the example above, if $Z_\theta^{(j)}$ are constructed in such a way that a better fit between the sample and $\mathbb{P}_\theta$ corresponds to a lower value, then having a lower rank on the original sample would imply a better estimate of the parameter. Therefore, the key idea is that a θ that minimizes $\mathcal{R}_\theta^{(m)}$ will be a good estimation, so a point estimate based on the resampling and ranking framamework can be defined as

$$\hat{\theta} \in \operatorname*{argmin}_{\theta\in\Theta} \mathcal{R}_\theta^{(m)} \tag{III.10}$$

However, there are some problems with this approach. First, the rank depends on the resamplings, which are random variables that would introduce a lot of noise into the value of the ranking function at each parameter, therefore the seed ξ needs to be fixed order for the minimum to be well defined.

The other problem is that if the seed ξ is fixed for the reference variable as well as the resamplings, and the distribution is parameterized reasonably, the value of the normalized rank given parameter θ will be a piecewise constant function, on which it would be difficult to optimize using gradient descent methods. Hence, the concept of *smoothed rank* was introduced (see Figure 2). It interpolates using the ordered version of $\{Z_\theta^{(i)}\}_{i\neq 0}$ at each point $\theta \in \Theta$, and in order to have a slope where the reference variable is the highest, the user can give any monotone function with some conditions.

Definition III.4. *[17, 4.3] Let $Y_\theta^{(1)} \leq ... \leq Y_\theta^{(m-1)}$ denote the pointwise ordered version of $\{Z_\theta^{(i)}\}_{i\neq 0}$. (For a formal definition of pointwise ordering, see III.13) Then the smoothed rank of $z \in \mathbb{R}$ is defined as:*

$$\tilde{\mathcal{R}}_{\theta,\xi}^{(m)}(z) = \begin{cases} \frac{1}{m}\left(1 + \frac{z}{Y_\theta^{(1)}}\right) & \text{if } z < Y_\theta^{(1)} \\ \frac{1}{m}\left(k+1+\frac{z-Y_\theta^{(k)}}{Y_\theta^{(k+1)}-Y_\theta^{(k)}}\right) & \text{if } Y_\theta^{(k)} \leq z < Y_\theta^{(k+1)} \\ \frac{1}{m}\left(m+\tau\left(z, Y_\theta^{(m-1)}\right)\right) & \text{if } Y_\theta^{(m-1)} \leq z \end{cases} \tag{III.11}$$

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 $Z_\theta^{(1)}$ respectively, 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 the corresponding regions 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 $\tau(z,y) = \frac{z^2}{y^2} - 1$

The definition is given for any $z \in \mathbb{R}$ (as this definition will be important in later sections), and from it the *smoothed rank of the reference variable* $\tilde{\mathcal{R}}_{\theta,\xi}^{(m)}$ can be defined as

$$\tilde{\mathcal{R}}_{\theta,\xi}^{(m)} = \tilde{\mathcal{R}}_{\theta,\xi}^{(m)}(Z_{\theta,\xi}^{(0)}) \tag{III.12}$$

The continuity of the construction above is entirely dependent on the continuity of the reference variables, as we will show next.

Lemma III.5. *Let (Θ, d) be a metric space and $Z^{(k)} : \Theta \to \mathbb{R}$ $(k \in [m])$ continuous functions. Denote their pointwise ordered version with $Z_*^{(i)}$:*

$$Z_*^{(i)}(\theta) = \min_{j\in[m]} \left\{ Z^{(j)}(\theta) \mid \#\left\{k \mid Z^{(j)}(\theta) \geq Z^{(k)}(\theta)\right\} \geq i \right\} \tag{III.13}$$

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

Corollary III.6. *If $Z^{(i)}$ are continuous in the parameter space Θ and $\mathbb{P}_\theta(Z_\theta^{(i)} = Z_\theta^{(j)}) = 0$ if $i \neq j$ for every $\theta \in \Theta$, then $\tilde{\mathcal{R}}_\theta$ is continuous with probability one.*

Proof. The smoothed rank $\tilde{\mathcal{R}}$ is constructed from elementary operations of $Z_\theta^{(i)}$ and $Y_\theta^{(i)}$, both of which are continuous. (The restrictions on τ ensure that the smoothed rank will be continuous in the corresponding region as well.) □

Now that the continuity of $\tilde{\mathcal{R}}$ is ensured, stepwise optimization techniques can be used to find its minimum. Some experiments with well-known stochastic optimization algorithms and their results can be found later in Section V, but first we discuss some theoretical results about the asymptotic properties of the *normalized rank* and the *smoothed rank*.

IV. Asymptotic Behavior

An interesting question that can be asked is what happens if the number of resamplings (m) or in case of an i.i.d. sample, the number of elements in each sample (n) is increased and either of them tends to infinity.

A. Increasing the Number of Resamplings

First, the asymptotics in $m \to \infty$ will be discussed. Since the range of the ranking function is dependent on m, the *normalized rank* $\mathcal{R}_\theta^{(m)}$ will be used for this analysis, 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.

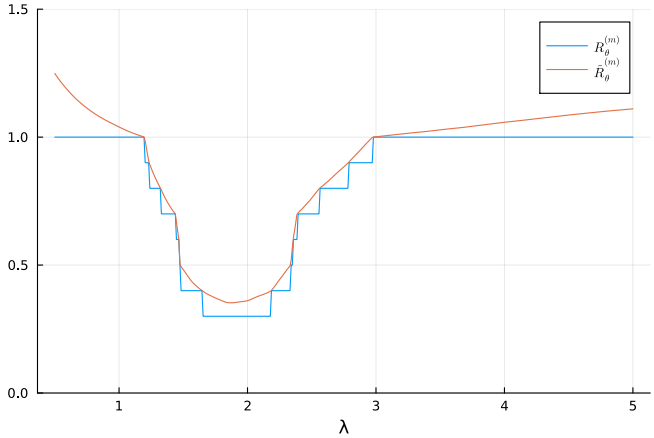


Fig. 2: The normalized rank and smoothed rank (for a fix seed) of a sample from an exponential distribution with parameter 2 using MMD based reference variables using the RBF kernel with $\sigma = 1$. ($n = 50$, $m = 10$, $\tau(z,y) = \frac{z}{y} - 1$)

In order to make the notation and the analysis clearer, the normalized rank of any $z \in \mathbb{R}$ can be defined as

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

The function $\mathcal{R}_\theta^{(m)}(z)$ can then be rewritten in the following form, so that it resembles an empirical cumulative distribution function a bit more:

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

From which, since $\{Z_\theta^{(i)}\}_{i\neq 0}$ are i.i.d., the law of large numbers can be applied in order to obtain

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

with probability one for any $z \in \mathbb{R}$, where $F_{Z_\theta^{(1)}}$ denotes the cumulative distribution function of $Z_\theta^{(1)}$.

First we assume that the original sample, $S^{(0)}$ is given and fix, but it will later be showed that this assumption can be relaxed.

Corollary IV.1. *If $Z_\theta^{(0)}$ is a deterministic reference variable (e.g. the seed is fixed or loglikelihood-based), then using the substitution $z = Z_\theta^{(0)}$, it holds with probability one that*

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

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. This is illustrated on Figure **??**, where the rank was calculated for a fix sample using an MMD-based reference variable with a fixed seed. We can see that the functions get more and more smooth as we get better and better approximations of the CDF.

If however, $Z_\theta^{(0)}$ is a random variable (for example an MMD-based reference variable is used), then because it is independent from $Z_\theta^{(1)}, ..., Z_\theta^{(m-1)}$, similarly to Corollary IV.1 there will be a convergence. Since the convergence holds for all $z \in \mathbb{R}$, even if both $\mathcal{R}_\theta^{(m)}$ and $Z_\theta^{(0)}$ are random variables, there will be a convergence with probability one:

Corollary IV.2. *If $Z_\theta^{(0)}$ is a randomized reference variable, then with probability one*

$$\lim_{m\to\infty} \mathcal{R}_\theta^{(m)} = F_{Z_\theta^{(1)}}\left(Z_\theta^{(0)}\right) \tag{IV.5}$$

Remark. It is important to emphasise that the rank at a fix parameter will be still be a random variable as $m \to \infty$, the corollaries above only describe its limiting distribution. However, if the reference variable is deterministic for $Z_\theta^{(0)}$ (for example by fixing the seed), then even if $Z_\theta^{(1)}, ..., Z_\theta^{(m-1)}$ are randomized, the limit is deterministic. (i.e. if we condition on the reference variable, then the limit is deterministic.)

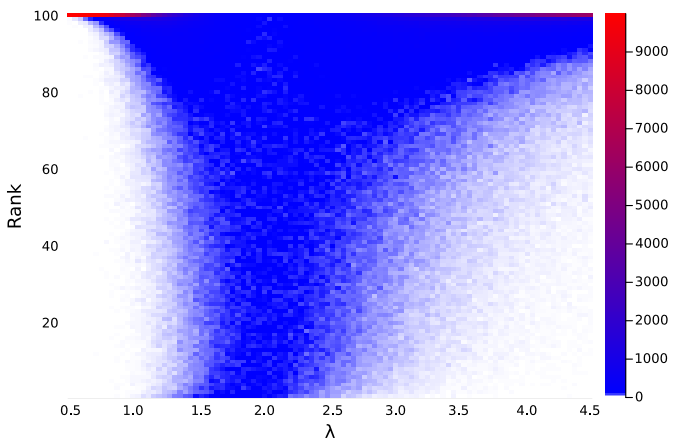


Fig. 3: The rank of a sample from an exponential distribution with parameter 2 in 10000 independent simulations at each parameter using MMD-based reference variables

We can also see that the randomness in the original sample can be incorporated into the randomness of the reference varaiable, and therefore Corollary IV.2 holds even if $S^{(0)}$ is not assumed to be fixed.

We can notice that the normalized rank $\mathcal{R}_\theta^{(m)}(z)$ almost corresponds to the empirical CDF of $\{Z_\theta^{(i)}\}_{i\neq 0}$ at point $z \in \mathbb{R}$, i.e. it can be rewritten as

$$\mathcal{R}_\theta^{(m)}(z) = \frac{1}{m} + \frac{m-1}{m} F_{m-1}(z) \tag{IV.6}$$

where $F_{m-1}(z)$ denotes the empirical cumulative distribution function of $\{Z_\theta^{(i)}\}_{i\neq 0}$ for a fix θ.

Remark. 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 $[0, 1]$ without conditioning on the original sample $S^{(0)}$. This is because substituting any random variable into its own CDF yields a uniform distribution.

Figure 3 illustrates this by showing that the distribution of the rank approaches a discrete uniform distribution as we get closer to the true parameter. In this simulation $S^{(0)}$ (of size $n = 50$) was redrawn every time from an exponential distribution with parameter 2 and was then ranked using an MMD-based reference variable and $m = 100$. There were 10000 trials at each parameter, and the number of times each rank is achieved is then visualised on a histogram (vertically) for each parameter. We can see that as we get closer to the true parameter $\lambda = 2$, the rank gets more and more uniformly distributed.

In order to give an upper give an upper bound for the rate of convergence in Corollary IV.2, the Dvoretzky-Kiefer-Wolfowitz inequality can be used.

Theorem IV.3 (Dvoretzky-Kiefer-Wolfowitz)**.** *[18] Let F_n denote the empirical CDF of n i.i.d. random variables with*

CDF F. Then for any $\lambda \in \mathbb{R}$ it holds that

$$\mathbb{P}\left(\sqrt{n}\sup_t |F_n(t) - F(t)| > \lambda\right) \leq 2\exp(-2\lambda^2) \quad \text{(IV.7)}$$

By rearranging the terms and using $\lambda = \sqrt{n}\varepsilon$ we obtain

$$\mathbb{P}\left(\sup_t |F_n(t) - F(t)| > \varepsilon\right) \leq 2\exp(-2n\varepsilon^2) \quad \text{(IV.8)}$$

This inequality can be used to give an upper bound for the probability of the difference between the normalized rank and the CDF being greater than some $\varepsilon > 0$ for any fixed θ:

Proposition IV.4. *Let $\mathcal{R}_\theta^{(m)}$ denote the normalized rank of the reference variable $Z_\theta^{(0)}$ and $F_{Z_\theta^{(1)}}$ the CDF of the alternative samples for the reference variable at parameter θ. Then the following inequality holds:*

$$\mathbb{P}\left(\left|\mathcal{R}_\theta^{(m)} - F_{Z_\theta^{(1)}}\left(Z_\theta^{(0)}\right)\right| > \varepsilon\right) \leq 2\exp\left(-2m\varepsilon^2 + 4\varepsilon\right)$$

Proof.

$$\begin{aligned}
&\mathbb{P}\left(\left|\mathcal{R}_\theta^{(m)} - F_{Z_\theta^{(1)}}\left(Z_\theta^{(0)}\right)\right| > \varepsilon\right) \leq \\
&\leq \mathbb{P}\left(\sup_z \left|\mathcal{R}_\theta^{(m)}(z) - F_{Z_\theta^{(1)}}(z)\right| > \varepsilon\right) \\
&= \mathbb{P}\left(\sup_z \left|\frac{1}{m} + \frac{m-1}{m}F_{m-1}(z) - F_{Z_\theta^{(1)}}(z)\right| > \varepsilon\right) \\
&\leq \mathbb{P}\left(\sup_z \frac{m-1}{m}\left|\left(F_{m-1}(z) - F_{Z_\theta^{(1)}}(z)\right)\right| + \frac{1}{m} > \varepsilon\right) \\
&= \mathbb{P}\left(\sup_z \left|\left(F_{m-1}(z) - F_{Z_\theta^{(1)}}(z)\right)\right| > \frac{m}{m-1}\left(\varepsilon - \frac{1}{m}\right)\right) \\
&\leq 2\exp\left(-2(m-1)\left(\frac{m}{m-1}\left(\varepsilon - \frac{1}{m}\right)\right)^2\right) \\
&= 2\exp\left(-2\frac{m^2}{m-1}\left(\varepsilon - \frac{1}{m}\right)^2\right) \\
&= 2\exp\left(-2\frac{1}{m-1}\left(m\varepsilon - 1\right)^2\right) \\
&\leq 2\exp\left(-2\frac{1}{m}\left(m^2\varepsilon^2 - 2m\varepsilon + 1\right)\right) \\
&\leq 2\exp\left(-2m\varepsilon^2 + 4\varepsilon\right)
\end{aligned}$$

□

We can see that for any fix $\varepsilon > 0$, an upper bound can be given across all possible reference variables for the probability of $\left|\mathcal{R}_\theta^{(m)} - F_{Z_\theta^{(1)}}\left(Z_\theta^{(0)}\right)\right| > \varepsilon$ at each point θ, which tends to zero as $m \to \infty$.

Next, we will apply Corollary IV.1 to the smoothed rank (Definition III.4). Since the ranking function will be piecewise constant only when the seed is fixed (otherwise there is noise present), we will restrict our discussion only to this case.

Remark. Before moving on to the theorem and its proof it is important to discuss and clarify the notion of *fixing the seed* for the reference variables and the subsamplings as $m \to \infty$.

We make the assumption that the seeds are contained in an infinite vector $\xi = (\xi_0, \xi_1, \xi_2, ...)$ where ξ_0 is used to calculate $Z_{\theta,\xi}^{(0)}$, ξ_1 is used for $Z_{\theta,\xi}^{(1)}$, and so on and $\{\xi_i\}_{i>0}$ were sampled i.i.d. from the same distribution Q. This means that the distribution of $\{Z_{\theta,\xi}^{(i)}\}_{i>0}$ will be as if they were sampled i.i.d., and therefore

$$\lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}_{\{Z_{\theta,\xi}^{(i)} < z\}} = \lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}_{\{Z_\theta^{(i)} < z\}} \quad \text{(IV.9)}$$

with probability 1.

Theorem IV.5. *Let $\tilde{\mathcal{R}}_{\theta,\xi}^{(m)}(z)$ denote the smoothed rank of $z \in \mathbb{R}$ for a fixed seed ξ, and $Z_\theta^{(1)}$ denote the alternative reference variable at point θ without fixing the seed. Then*

$$\lim_{m\to\infty}\tilde{\mathcal{R}}_{\theta,\xi}^{(m)}(z) = \mathbb{P}\left(Z_\theta^{(1)} < z\right) = F_{Z_\theta^{(1)}}(z) \quad \text{(IV.10)}$$

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

$$\lim_{m\to\infty}\tilde{\mathcal{R}}_{\theta,\xi}^{(m)} = F_{Z_\theta^{(1)}}\left(Z_{\theta,\xi}^{(0)}\right) \quad \text{(IV.11)}$$

with probability one for any fixed seed ξ.

Remark. Similarly to the normalized rank, under the null hypothesis $\mathbb{P}_\theta = \mathbb{P}_{\theta^*}$, if $Z_\theta^{(i)}$ are continuous random variables, then $\lim_{m\to\infty}\tilde{\mathcal{R}}_{\theta,\xi}^{(m)} = F_{Z_\theta^{(0)}}\left(Z_\theta^{(0)}\right)$ is uniformly distributed over $[0,1]$ if we randomize both the seed and the original sample.

B. Uniform Convergence

In the previous section we have shown that there is a pointwise convergence at each parameter θ for the ranking function. Since this convergence shows similarities to the Glivenco-Cantelli Theorem [15, Proposition 3.24], the question of whether it is possible to guarantee uniform convergence across the parameter spaces can be asked. It is important to emphasise that the Glivenco-Cantelli Theorem provides uniform convergence over the *sample space*, but we want it over the *parameter space*.

In this section we will use the tools of VC-theory, and a generalisation of the Glivenco-Cantelli Theorem in order to state a sufficient condition for uniform convergence, that depends only on the type of reference variable used. Then we will show applications of this theorem, applying it for the previously defined reference variables, and even showing a uniform converge across the sample space.

We start by stating the *UCG (Uniform Glivenco-Cantelli)* property, which defines what is meant by *uniform convergence*. This definition would ensure that if the ranking function is written in place of the empirical mean of f (see in the definition), then for any distribution of the resamplings (which is based on the hypotethised distribution), there would be a *uniform convergence*.

Definition IV.7. *[19, Definition 3.23] Let $\mathcal{X}$ be a standard Borel space and $M_\mathcal{X}$ be the set of all probability distributions over $\mathcal{X}$. A set of $\mathcal{X} \to \mathbb{R}$ measurable functions $\mathcal{H}$ is uniform Glivenco-Cantelli (UGC) if for every $\varepsilon > 0$,*

$$\lim_{l\to\infty}\sup_{\mu\in M_\mathcal{X}}\mathbb{P}\left(\sup_{m\geq l}\sup_{f\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^{m}f(x_i) - \int_\mathcal{X} f(x)\,d\mu\right| \geq \varepsilon\right) = 0 \quad \text{(IV.12)}$$

where $\{x_i\}$ are sampled i.i.d. from the distribution μ.

Next, we state the definition of VC-dimension. This concept is widely used in statistical learning theory to measure the *capacity* of function classes [20][21]. In this case, it will be used to give an equivalent condition for the UGC property.

Definition IV.8. *[21, Definition 4.1] Let $\mathcal{X}$ be a measurable space and $\mathcal{H}$ be a collection of $\mathcal{X} \to \{0,1\}$ functions. We say that $\mathcal{H}$ shatters a set T if for every $U \subset T$ there exists an $f_U \in \mathcal{H}$ such that $f_U(x) = 1$ for all $x \in U$ and $f_U(x) = 0$ for all $x \in T \setminus U$. The VC dimension of $\mathcal{H}$ is defined as the maximum cardinality of a set that $\mathcal{H}$ can shatter.*

As it can be seen in the definition above, the concept of VC-dimension was first introduced for binary classification problems. There are however different generalisations of it for regression problems, one being the fat shattering dimension defined bellow as V_γ-dimension:

Definition IV.9. *[19, Definition 3.24] Let $\mathcal{H}$ be a set of $\mathcal{X} \to [0,1]$ functions and $\gamma > 0$. We say that $A \subset \mathcal{X}$ is V_γ shattered by $\mathcal{H}$ if $\exists \alpha \in \mathbb{R}$ such that for every $E \subset A$ there exists an $f_E \in \mathcal{H}$ s.t. $f_E(x) \leq \alpha - \gamma$ for every $x \in A \setminus E$ and $f_E(x) \geq \alpha + \gamma$ for every $x \in E$. The V_γ-dimension of $\mathcal{H}$, $V_\gamma(\mathcal{H})$ is the maximal cardinality of a set A that can be V_γ shattered by $\mathcal{H}$.*

Remark. The definition of V_γ-dimension can also be applied to binary classificaton problems. In this case if $\mathcal{H}$ is a set of $\mathcal{X} \to \{0,1\}$ functions, then its V_γ dimension is equal to its VC-dimension if $\gamma \leq 1/2$, and therefore $V_\gamma(\mathcal{H}) \leq \text{VC}(\mathcal{H})$ for every γ.

Now we state the theorem (Which is a generalisation of the Glivenco-Cantelli Theorem), that was the motivation for introducing the capacity measures previously.

Theorem IV.10. *[19, Theorem 3.25] Let $\mathcal{H}$ be a set of $\mathcal{X} \to [0,1]$ functions. Then $\mathcal{H}$ is UGC if and only if the $V_\gamma(\mathcal{H})$ is finite for every $\gamma > 0$*

This theorem can also be applied to a set of binary functions as well, since $V_\gamma(\mathcal{H}) \leq \text{VC}_(\mathcal{H})$:

Corollary IV.11. *If $\mathcal{H}$ is a set of $\mathcal{X} \to \{0,1\}$ functions and $VC(\mathcal{H}) < \infty$, then Θ is UCG.*

The motivation for stating Corollary IV.11 is that the normalized rank is a mean of indicator functions (indicating whether the reference variable is greater than an alternative variable), and therefore we need to have a restriction on the VC-dimension of these indicator functions. However, for the theorem to be applicable, we need to have the condition on the reference variable, not the indicator functions. Luckily the concept of *pseudo-dimension* can be used in order to ensure this:

Definition IV.12. *[21, Definition 4.2] Let $\mathcal{G}$ be a set of real valued functions on $\mathcal{X}$ and $S = \{x_1, ..., x_m\} \subset \mathcal{X}$. Then S is pseudo-shattered by $\mathcal{G}$ if there exists $z = (z_1, ..., z_m) \in \mathbb{R}^m$ such that for every $E \subset S$ there exists an $f_E \in \mathcal{H}$ for which $f_E(x_i) > z_i$ if $x_i \in E$ and $f_E(x_i) \leq z_i$ if $x_i \in S \setminus E$. We say that z is a witness to the pseudo-shattering. The pseudo-dimension of $\mathcal{G}$ denoted as Pdim$(\mathcal{G})$ is the maximal cardinality of a set S that can be pseudo-shattered by $\mathcal{G}$.*

Lemma IV.13. *[21, Lemma 4.1] If $\mathcal{G}$ is a set of $\mathcal{X} \to \mathbb{R}$ functions and a set of $\mathcal{X} \times \mathbb{R} \to \{0,1\}$ functions $\mathcal{H}$ is defined as*

$$\mathcal{H} = \{f((x,z)) = \mathbb{I}_{\{g(x) > z\}} | g \in \mathcal{G}\} \tag{IV.13}$$

then VC$(\mathcal{H})$ = Pdim$(\mathcal{G})$

This means that we can use a *pseudo-dimension* based condition for the *UCG* property across the parameter space Θ, as well as all possible seeds $\xi \in \mathcal{Q}$:

Theorem IV.14. *Let $S^{(0)}$ denote a fixed original sample, $Z^{(0)}_{\theta,\xi}$ its reference variable calculated using seed ξ and $\mathcal{R}^{(m)}_{\theta,\xi}$ be its rank. If there exists a set of $\Theta \to \mathbb{R}$ functions $\mathcal{G}$ such that $Z^{(0)}_{\theta,\xi} \in \mathcal{G}$ for every seed $\xi \in \mathcal{Q}$, and Pdim$(\mathcal{G}) < \infty$, then*

$$\lim_{l \to \infty} \sup_{\theta \in \Theta} \mathbb{P}\left(\sup_{m \geq l} \sup_{\xi \in \mathcal{Q}} \left| \mathcal{R}^{(m)}_{\theta,\xi} - F_{Z^{(1)}_\theta}\left(Z^{(0)}_{\theta,\xi}\right) \right| \geq \varepsilon \middle| S^{(0)} \right) = 0 \tag{IV.14}$$

Proof. Let's define the variables according to the notations of Definition IV.7:

If $Z^{(1)}_\theta$ is a random variable, then the distribution of $x_1 = (\theta, Z^{(1)}_\theta)$ denoted as μ_θ is Borel-measurable on $\mathcal{X} = \Theta \times \mathbb{R}$ for any fix θ. (Note that the first coordinate is just the point measure at θ.) Therefore $\{\mu_\theta\}_{\theta \in \Theta} = M' \subset M_{\mathcal{X}}$. Next, the function class $\mathcal{H}$ is defined as

$$\mathcal{H} = \left\{ f_g(x) = f_g((\theta, Z^{(1)}_\theta)) = \mathbb{I}_{\{Z^{(1)}_\theta < g(\theta)\}} \middle| g \in \mathcal{G} \right\} \tag{IV.15}$$

Since $\mathcal{H}$ is a class of binary functions, according to Theorem IV.10, there will be a convergence on all Borel probability measures if VC$(\mathcal{H}) < \infty$, which is equivalent to Pdim$(\mathcal{G}) < \infty$. This results in a uniform convergence over $\mathcal{G}$ of $\frac{1}{m}\sum_{i=1}^m \mathbb{I}_{\{z_i < g(\theta_i)\}}$ to $\mathbb{P}(z_i < g(\theta_i))$ for all $\mu \in M_{\mathcal{X}}$. (Here the samples $\{x_i = (\theta_i, z_i)\}_{i=1}^m$ could come from any $\mu \in M_{\mathcal{X}}$) Therefore the convergence also on M'. For any $\mu_\theta \in M'$ the set of samples $\{(x_1, \theta_1), ..., (x_m, \theta_m)\}$ can be written as $\{(x_1, \theta), ...(x_m, \theta)\}$ and therefore

$$\frac{1}{m}\sum_{i=1}^m \mathbb{I}_{\{z_i < g(\theta_i)\}} = \frac{1}{m}\sum_{i=1}^m \mathbb{I}_{\{Z^{(i)}_\theta < g(\theta)\}} = F_m(g(\theta)) \tag{IV.16}$$

and

$$\mathbb{P}(z_i < g(\theta_i)) = \mathbb{P}\left(Z^{(i)}_\theta < g(\theta)\right) = F_{Z^{(1)}_\theta}(g(\theta)) \tag{IV.17}$$

where F_m denotes the empirical CDF of $\{Z^{(i)}_\theta\}_{i=1}^m$. From IV.6 it can be seen that the convergence of the empirical CDF is equivalent to the convergence of the rank at each point, and the convergence speed is only dependent on m. Therefore $\mathcal{R}^{(m)}_\theta(g(\theta))$ converges to $F_{Z^{(1)}_\theta}(g(\theta))$ uniformly on θ for every $g \in \mathcal{G}$. Since $Z^{(0)}_{\theta,\xi} \in \mathcal{G}$ for every fix seed ξ, the convergence should also hold for $\mathcal{R}^{(m)}_{\theta,\xi} = \mathcal{R}^{(m)}_\theta(Z^{(0)}_{\theta,\xi})$. □

In order to be able to apply the theorem above to some reference variables, we will give some basic arithmetic properties that can be used to prove that a function class has a finite *pseudo-dimension*.

Theorem IV.15. *[22, Theorem 11.4] If $\mathcal{H}$ is a vector space of real-valued functions, then Pdim$(\mathcal{H}) = \dim(\mathcal{H})$*

Corollary IV.16. *[22, Corollary 11.5] If $\mathcal{H} \subset \mathcal{F}$ where $\mathcal{F}$ is a vector space, then Pdim$(\mathcal{H}) \leq \dim(\mathcal{F})$*

Lemma IV.17. *[22, Theorem 11.3] Let $\mathcal{H}$ be a class of $\mathcal{X} \to \mathbb{R}$ functions, and $g : \mathbb{R} \to \mathbb{R}$ a non-decreasing function. Then for the function class $\mathcal{G} = \{g(f(x)) | f \in \mathcal{H}\}$ it holds that Pdim$(\mathcal{G}) \leq$ Pdim$(\mathcal{H})$*

Lemma IV.18. *Let $\mathcal{H}$ be a set of $\mathcal{X} \to \mathbb{R}$ functions and $g : \Psi \to \mathcal{X}$ be any function. Then for the pseudo-dimension of $\mathcal{G} = \{h(g(\psi)) | h \in \mathcal{H}\}$ (where $\mathcal{G}$ is a set of $\Psi \to \mathbb{R}$ functions) it holds that Pdim$(\mathcal{G}) \leq$ Pdim$(\mathcal{H})$.*

Proof. Let $X = \{\psi_1, ..., \psi_n\} \subset \Psi$ such that it can be pseudo-shattered by $\mathcal{G}$. First we prove that the set $X' = \{g(\psi_1), ..., g(\psi_n)\} \subset \Theta_1$ has cardinality n, i.e. $g(\psi_i) \neq g(\psi_j)$ if $i \neq j$.

Assume that there are two indices $i \neq j$ for which $g(\psi_i) = g(\psi_j)$, denote their *witnesses* by z_i and z_j, and let $E_1 = \{\psi_i\}$ and $E_2 = \{\psi_j\}$. This would mean that there exist functions $f_{E_1}, f_{E_2} \in \mathcal{H}$ for which

$$\begin{aligned} z_i < f_{E_1}(g(\psi_i)) = f_{E_1}(g(\psi_j)) \leq z_j \\ z_j < f_{E_2}(g(\psi_j)) = f_{E_1}(g(\psi_i)) \leq z_i \end{aligned} \tag{IV.18}$$

which is a contradiction, and therefore $|X'| = n$.

The other required condition is X' to be pseudo-shattered by $\mathcal{H}$. This holds, because every $h \circ g$ function from $\mathcal{G}$ creates the same partition on X as h does on X', and therefore Pdim$(\mathcal{G}) \leq$ Pdim$(\mathcal{H})$. □

First we can notice that if the reference variable is not randomized, then the uniform convergence property holds. In this case it can be assumed that $\mathcal{G}$ contains only one function, namely the reference variable, and therefore $|\mathcal{G}| = 1$, resulting in Pdim$(\mathcal{G}) = 0$. Similarly, the seed being fixed, or the seed space $\mathcal{Q}$ having a finite cardinality results in the pseudo-dimension of $\mathcal{G}$ being finite, meaning a uniform convergence. Note that (similarly to the case of pointwise convergence) the seed is only needed to be fixed for the reference variable, and not for the alternative reference variables in order for the UCG property to hold.

However, the results above only hold if we assume that the original sample $S^{(0)} \in \mathcal{X}^m$ is fixed. Fortunately, this assumption can be relaxed if the reference variable is in $\mathcal{G}$ for every possible seed and original sample i.e. $Z^{(0)}_{\theta,\xi} : \theta \to \mathbb{R} \in \mathcal{G}$ for every $(S^{(0)}, \xi) \in \mathcal{X}^m \times \mathcal{Q}$ and Pdim$(\mathcal{G}) < \infty$.

This relaxation will be showcased by the following proposition, where the uniform convergence across all original samples is proved for Loglikelihood-based reference variables from the exponential distribution family.

Proposition IV.19. *Let $Z^{(0)}_\theta$ be a Loglikelihood-based reference variable using the Likelihood-function of a distribution from a d-dimensional exponential family, i.e. the reference variable of the original sample $S^{(0)} = (x_1, ..., x_n)$ at parameter $\theta \in \mathbb{R}^p$ can be expressed in the form*

$$Z^{(0)}_\theta = \left\| \nabla_\theta \log L(\theta, S^{(0)}) \right\|^2 \tag{IV.19}$$

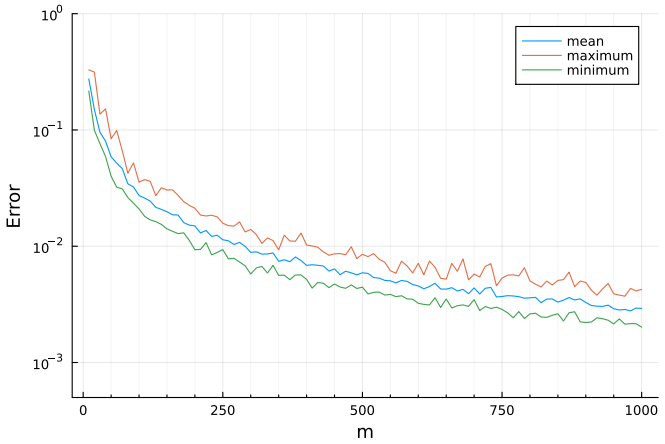


Fig. 4: The mean, maximum and minimum of 50 independent trials for the maximum of $(\mathcal{R}^{(m)}_\theta - F_{Z^{(1)}_\theta}(Z^{(0)}_\theta))^2$ over the parameter space with increasing m illustrates the uniform convergence when the loglikelihood-based reference variable is used. The sample of size 50 was drawn from an exponential distribution with parameter 1. The maximum was taken over the discretized parameter space $[0.01, 3]$ divided into 10000 pieces, with the alternative reference variables being redrawn at every step independently.

where L denotes the likelihood of θ given $S^{(0)}$:

$$L\left(\theta, S^{(0)}\right) = \prod_{i=1}^{n} h(x_i) \exp\left(\theta^{\mathrm{T}} T(x_i) - b(\theta)\right) \tag{IV.20}$$

Then there is a uniform convergence of the rank of the reference variable to its theoretical limit across the sample space, as well as the parameter space:

$$\lim_{l \to \infty} \sup_{\theta \in \Theta} \mathbb{P}\left(\sup_{m \geq l} \sup_{S^{(0)} \subset \mathbb{R}^d} \left| \mathcal{R}^{(m)}_\theta - F_{Z^{(1)}_\theta}\left(Z^{(0)}_\theta\right) \right| \geq \varepsilon \right) = 0 \tag{IV.21}$$

Proof. The value of the reference variable can be rewritten as

$$\begin{aligned} Z^{(0)}_\theta &= \left\| \nabla_\theta \mathcal{L}(\theta, S^{(0)}) \right\|^2 \\ &= \left\| \nabla_\theta \left(\sum_{i=1}^{n} \log\left(h(x) \exp\left(\theta^{\mathrm{T}} T(x_i) - b(\theta)\right)\right) \right) \right\|^2 \\ &= \left\| \nabla_\theta \left(\sum_{i=1}^{n} \log(h(x_i)) + \sum_{i=1}^{n} \theta^T T(x_i) - n b(\theta) \right) \right\|^2 \\ &= \left\| \sum_{i=1}^{n} T(x_i) - n \nabla_\theta b(\theta) \right\|^2 \end{aligned}$$

Here we can define $\xi = \sum_{i=1}^{n} T(x_i)$ and $\beta(\theta) = n \nabla_\theta b(\theta)$, and therefore

$$Z^{(0)}_{\theta,\xi} = ||\xi - \beta(\theta)||^2 \tag{IV.22}$$

It is important to note that only the value of $\xi \in \mathbb{R}^p$ is dependent on the original sample and it is independent of θ.

This means that all possible reference variables that could be assigned to $S^{(0)}$ at parameter θ are contained in the set

$$\mathcal{G} = \left\{ ||\xi - \beta(\theta)||^2 : \xi = \sum_{i=1}^{n} T(x_i) \Big| x_1, ..., x_n \in \mathbb{R}^d \right\} \tag{IV.23}$$

The goal is now to prove that $\text{Pdim}(\mathcal{G}) < \infty$, so that Theorem IV.14 can be applied in order to obtain the uniform convergence. To proceed with this, first we relax the condition on ξ, resulting in a larger function class:

$$\mathcal{G}' = \left\{ \|\xi - \beta(\theta)\|^2 \Big| \xi \in \mathbb{R}^p \right\} \tag{IV.24}$$

Since $\mathcal{G} \subset \mathcal{G}'$, it is enough to prove $\text{Pdim}(\mathcal{G}') < \infty$. From Lemma IV.18 it can be seen that it is enough to prove that the function class $\mathcal{G}'' = \{f_v(u) = \|u - v\|^2 | v \in \mathbb{R}^d\}$ has a finite pseudo-dimension. Every element of $\mathcal{G}''$ can be rewritten as

$$f_v(u) = \sum_{i=1}^{d} v_i^2 - 2v_i^{\mathrm{T}} u_i + u_i^2 = \sum_{i=1}^{d} v_i^2 + \sum_{i=1}^{d} u_i^2 - 2\sum_{i=1}^{d} v_i^{\mathrm{T}} u_i \tag{IV.25}$$

where the changing values of $\{v_i\}$ determine the function class. From this representation it can be seen that this function class is a subset of a $2d+1$-dimensional vector space $\mathcal{H}$:

$$\mathcal{H} = \left\{ h_{w,v,a}(u', u) = a + \sum_{i=1}^{d} w_i^{\mathrm{T}} u_i' + \sum_{i=1}^{d} v_i^{\mathrm{T}} u_i \right\} \tag{IV.26}$$

and therefore

$$\text{Pdim}(\mathcal{G}) \leq \text{Pdim}(\mathcal{G}') \leq \text{Pdim}(\mathcal{G}'') \leq \text{Pdim}(\mathcal{H}) \leq 2d+1 \tag{IV.27}$$

Meaning that Theorem IV.14 can be applied to function class $\mathcal{G}$, resulting in a uniform convergence across the sample space. □

From the previous example, we can see that Theorem IV.14 can be used to prove uniform convergence on Θ across all possible original samples. We prove in a similar manner the uniform convergence across all possible original samples for certain reference variables if the seed is fixed.

Theorem IV.20. *Let $S^{(0)}$ denote the original sample, $Z_{\theta,\xi}^{(0)}$ its reference variable calculated using a fixed seed ξ and let $\mathcal{R}_{\theta,\xi}^{(m)}$ be its rank. If there exists a set of $\Theta \to \mathbb{R}$ functions $\mathcal{G}$ such that $Z_{\theta,\xi}^{(0)} \in \mathcal{G}$ for every original sample $S^{(0)}$, and Pdim$(\mathcal{G}) < \infty$, then it holds that*

$$\lim_{l\to\infty} \sup_{\theta\in\Theta} \mathbb{P}\left(\sup_{m\geq l} \sup_{S^{(0)}\in\mathcal{X}^n} \left| \mathcal{R}_{\theta,\xi}^{(m)} - F_{Z_\theta^{(1)}}\left(Z_{\theta,\xi}^{(0)}\right) \right| \geq \varepsilon \Big| \xi \right) = 0 \tag{IV.28}$$

Proof. The proof is the same as for Theorem IV.14, since the only difference is between how we define the function class, and the only property used to define the uniform convergence is $\text{Pdim}(\mathcal{G}) < \infty$. □

We will use the Theorem IV.20 to prove uniform convergence across all original samples for MMD-based reference variables constructed using a finite-dimensional RKHS.

Proposition IV.21. *Let $Z_\theta^{(0)}$ be an MMD-based reference variable, that was constructed using a kernel k, for which its corresponding RKHS $\mathcal{H}$ is a finite-dimensional vector space. If the seed for calculating $Z_\theta^{(0)}$ is fixed, then there is a uniform convergence across all possible original samples $S^{(0)} \in \mathcal{X}^n$:*

$$\lim_{l\to\infty} \sup_{\theta\in\Theta} \mathbb{P}\left(\sup_{m\geq l} \sup_{S^{(0)}\in\mathcal{X}^n} \left| \mathcal{R}_{\theta,\xi}^{(m)} - F_{Z_\theta^{(1)}}\left(Z_{\theta,\xi}^{(0)}\right) \right| \geq \varepsilon \Big| \xi \right) = 0 \tag{IV.29}$$

Proof. Let $Z_{\theta,\xi}^{(0)} = \widehat{\text{MMD}}_{\mathcal{H}}^2[S^{(0)}, S_\xi^{(m)}(\theta)]$ where $S_\xi^{(m)} : \theta \to \mathcal{X}^n$ denotes the function that generates an alternative sample for the fix seed ξ. We will once again try to prove that

$$\text{Pdim}\left(\left\{ Z_{\theta,\xi}^{(0)} = \widehat{\text{MMD}}_{\mathcal{H}}^2[S^{(0)}, S_\xi^{(m)}(\theta)] \Big| S^{(0)} \in \mathcal{X}^n \right\} \right) < \infty \tag{IV.30}$$

First, we use Lemma IV.18 to see that by defining a set of $\mathcal{X}^n \to \mathbb{R}$ functions as

$$\mathcal{G} = \left\{ \widehat{\text{MMD}}_{\mathcal{H}}^2[S^{(0)}, S] \big| S^{(0)} \in \mathcal{X}^n \right\} \tag{IV.31}$$

we have

$$\text{Pdim}\left(\left\{ Z_{\theta,\xi}^{(0)} \Big| S^{(0)} \in \mathcal{X}^n \right\} \right) \leq \text{Pdim}\left(\mathcal{G}\right) \tag{IV.32}$$

Next, we write out a function from $\mathcal{G}$ in its full form where $\{x_i\}$ are the elements of $S^{(0)}$ that parameterize the function class and $\{y_j\}$ are the elements of S, the input of the function.

$$\begin{aligned} f_{S^{(0)}}(S) &= \widehat{\text{MMD}}_{\mathcal{H}}^2[S^{(0)}, S] \\ &= \frac{1}{n(n-1)} \sum_{i=1}^{n}\sum_{j=1}^{n} k(x_i, x_j) \\ &+ \frac{1}{n(n-1)} \sum_{i=1}^{n}\sum_{j=1}^{n} k(y_i, y_j) \\ &- \frac{2}{n^2} \sum_{i=1}^{n}\sum_{j=1}^{n} k(x_i, y_j) \end{aligned} \tag{IV.33}$$

Let $e_1, ..., e_d$ denote an orthonormal basis of $\mathcal{H}$. From the reproducing property of $(\mathcal{H}, k)$, for any $u, v \in \mathcal{H}$, $k(u, v)$ can be rewritten as

$$k(u, v) = \langle k_v, k_u \rangle = \langle \alpha_1 e_1 + \alpha_d e_d, \beta_1 e_1 + \beta_d e_d \rangle = \sum_{i=1}^{d} \alpha_i \beta_i$$

Now we can see that $\mathcal{G}$ is a subset of a finite-dimensional vector space: Let $k_{x_i} = \alpha_{i,1} e_1 + ... + \alpha_{i,d} e_d$ for every $x_i \in S^{(0)}$. (Notice that $\mathcal{G}$ is parameterized by $\alpha = (\alpha_{1,1}, ..., \alpha_{n,d})$, since $f_{S^{(0)}}$ depends on $S^{(0)}$ only through $k_{x_1}, ..., k_{x_n}$.)

Similarly, for a fix $S^{(0)}$ the value of $f_{S^{(0)}}(S)$ depends on S only through

$$\{k_{y_i}\}_{i=1}^n = \{(\beta_{i,1} e_1 + ... + \beta_{i,d} e_d)\}_{i=1}^n \tag{IV.34}$$

By introducing the function $\psi(S) = (\beta_{1,1}, ..., \beta_{n,d}) = \beta$, Lemma IV.18 can be used on $f_\alpha(\psi(S)) = f_\alpha(\beta)$ where

$$f_\alpha(\beta) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{l=1}^{d} \alpha_{i,l}\alpha_{j,l} + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{l=1}^{d} \beta_{i,l}\beta_{j,l} - \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{l=1}^{d} \alpha_{i,l}\beta_{j,l} \tag{IV.35}$$

Similarly to equation IV.26 we can see that this is a subset of a $2dn^2 + 1$ dimensional space, and therefore it has a finite pseudo-dimension. □

C. Increasing Sample Size

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

Definition IV.22. *We say that a reference variable is consistent, if it holds that*

$$\lim_{n\to\infty} Z_\theta^{(i)} = \begin{cases} 0 & \text{if } x_j \sim \mathbb{P}_\theta \text{ i.i.d.} \\ c \in \mathbb{R}_+ \cup \{\infty\} & \text{else} \end{cases} \tag{IV.36}$$

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

Proposition IV.23. *(Pointwise consistency) If $Z_\theta^{(i)}$ are consistent, then for any fix $\theta \in \Theta$, the normalized 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 $\left\{\frac{1}{m}, ..., \frac{m-1}{m}, 1\right\}$ if the original sample $S^{(0)}$ is considered random as well.

Next, we prove that the MMD based reference variables are consistent if they are constructed using a characteristic kernel. An example for increasing the size of the sample for a characteristic (in this case RBF) kernel is shown in Figure **??**. To show that characteristic kernels are consistent, first we need the law of large numbers for kernel mean embeddings.

Lemma IV.24. *(Law of large numbers for kernel mean embeddings) Let $\mathcal{H}$ be a real, separable 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 $n \to \infty$.*

Corollary IV.25. *If an MMD-based reference variable is constructed using a characteristic kernel, then it is consistent.*

V. Optimization

A. Algorithms

In this section we explore some stochastic approximation algorithms that can be used to find a point estimate for the minimizer $\hat{\theta}$ of the ranking function. It will be assumed that the distribution is parameterised by a vector space Θ, so that gradient descent-based algorithms can be be used (motivated by the widely used stochastic gradient descent algorithm [23]). Unfortunately however, the gradient of the rank cannot be calculated, since we assume that (both the original and the alternative) samples are generated by a black box $G : \Theta \times \mathcal{Q} \to \mathcal{X}$, and therefore we do not explicitly know how the rank depends on parameter θ. In order to overcome this problem, we will take inspiration from stochastic approximation algorithms. One such algorithm is the Kiefer-Wolfowitz algorithm [24], that can be used for one-dimensional parameter spaces. This algorithm works by taking a small step in each direction, and then estimating the gradient based on the value of the function in each direction:

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

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

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 $\sum_{n=0}^\infty \frac{\gamma_n^2}{\delta_n^2} < \infty$.

This idea can be extended to finite-dimensional vector spaces using the Simultaneous Perturbation Stochastic Approximation (SPSA) algorithm [25], which estimates the gradient locally by taking only one step in a random, as well as the exact opposite direciton.

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

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

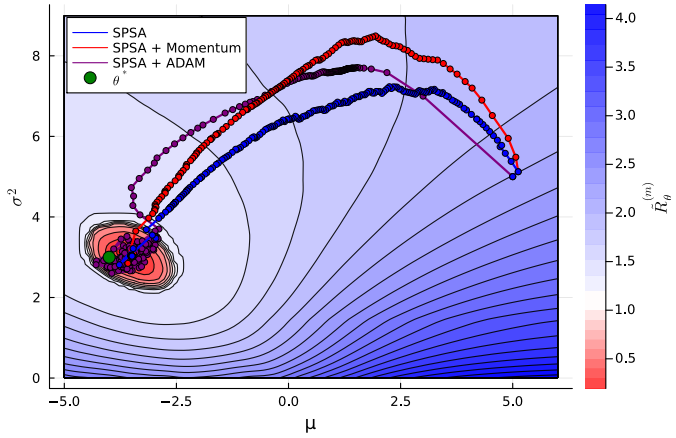
where $\theta_{n,k}$ denotes the kth coordinate of θ_n, and $\{\Delta_n\}$ are independent, symmetric, zero-mean vectors, for example Bernoulli trials with $\Delta_{n,k} = \pm 1$ with probability $\frac{1}{2}$ each.

We can further echance the algorithm above by adding a momentum component to it (making it similar to the heavy-ball method), or by making it similar to the popular ADAM algorithm [26, Algorithm 1].

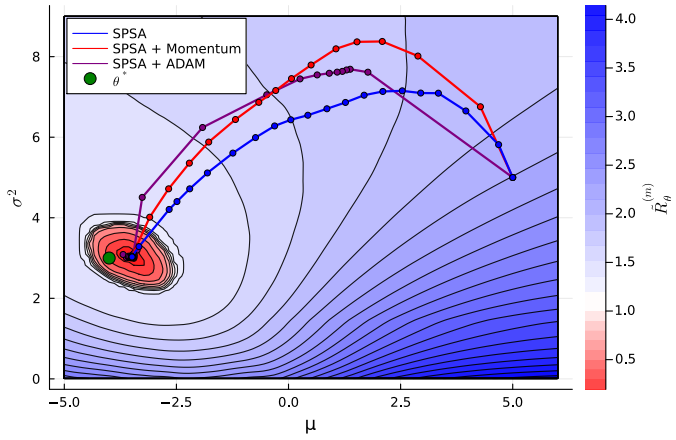
VI. Conclusion

There are many promising directions to continue on from here in both theory and practice. An interesting theoretical question is whether the theory of Rashomon sets [6] can be incorporated to the topic of ranking functions. A practical direction to further investigate is whether this framework can be applied to train or fine-tune more complex models (such as diffusion models for image generation), or to explore methods that can create region estimates.

In this work we have first established a framework for creating consistent, distribution-free confidence regions for generative models with exact coverage for finite samples. The asymptotic behavior of the ranking function (which is the basis of the tests used to construct the confidence regions) was investigated, giving an upper bound for the rate of convergence (as the number of resamplings is increased) at a fix parameter. We also gave theoretical guarantees under certain conditions



(a) Every step of each algorithm



(b) The average of every 10 steps for the algorithms

Fig. 5: SPSA based optimization methods 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.

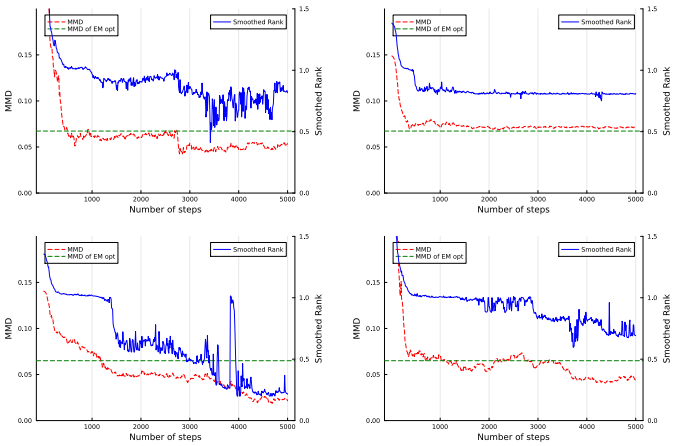


Fig. 6: Example runs for the resampling and ranking based SPSA-ADAM algorithm used for the estimation of a gaussian mixture model. The decrease in smoothed rank and MMD-distance from the original distribution over time can be observed, with the MMD of an EM-algorithm optimum displayed for reference.

for uniform convergence over the parameter space, collected applicable tools for the proposed condition and demonstrated usecases for it, proving a uniform convergence across the sample space for different types of reference variables. As for the limit of the rank in the size of the original sample, the definition of a consistent reference variable was introduced to give a sufficient condition for pointwise convergence, and it was proved that MMD-based reference variables constructed using characteristic kernels are consistent.

Methods for creating a point estimate using the introduced framework were also discussed. In this topic the continuity of the smoothed rank was proved and simulations were run for the parameter estimation using different optimization methods.

References

[1] R. J. Rossi, *Mathematical statistics: an introduction to likelihood based inference*. John Wiley & Sons, 2018.

[2] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[3] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, 2020.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[5] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, pp. 1065–1076, 1962.

[6] L. Semenova, C. Rudin, and R. Parr, "On the existence of simpler machine learning models," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1827–1858.

[7] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[8] C. Rudin, C. Zhong, L. Semenova, M. Seltzer, R. Parr, J. Liu, S. Katta, J. Donnelly, H. Chen, and Z. Boner, "Amazing things come from having many good models," *arXiv preprint arXiv:2407.04846*, 2024.

[9] B. C. Csáji, M. C. Campi, and E. Weyer, "Sign-perturbed sums: A new system identification approach for constructing exact non-asymptotic confidence regions in linear regression models," *IEEE Transactions on Signal Processing*, vol. 63, no. 1, pp. 169–181, 2014.

[10] 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*, pp. 860–865, 2022.

[11] B. C. Csáji and K. B. Kis, "Distribution-free uncertainty quantification for kernel methods by gradient perturbations," *Machine Learning*, vol. 108, no. 8, pp. 1677–1699, 2019.

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

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

[14] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf, "Kernel measures of conditional dependence," *Advances in Neural Information Processing Systems*, vol. 20, 2007.

[15] O. Kallenberg, *Foundations of modern probability*. Springer, 1997.

[16] L. Devroye, "Nonuniform random variate generation," *Handbooks in Operations Research and Management Science*, pp. 83–121, 2006.

[17] A. Jung, "Hypothesis test based estimation," *HUN-REN SZTAKI Technical Report*, 2024.

[18] P. Massart, "The tight constant in the dvoretzky-kiefer-wolfowitz inequality," *The Annals of Probability*, pp. 1269 – 1283, 1990.

[19] F. Cucker and D. X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*, ser. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2007.

[20] N. V. Vladimir and V. Vapnik, "Statistical learning theory," *Xu JH and Zhang XG. translation. Beijing: Publishing House of Electronics Industry, 2O04*, vol. 1, 1998.

[21] M. Vidyasagar, *Learning and generalisation: with applications to neural networks*. Springer Science & Business Media, 2013.
[22] M. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations*, 1st ed. Cambridge University Press, 2009.
[23] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, 2010 Keynote, Invited and Contributed Papers*. Springer, 2010, pp. 177–186.
[24] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *The Annals of Mathematical Statistics*, 1952.
[25] J. Spall, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Transactions on Automatic Control*, vol. 37, no. 3, pp. 332–341, 1992.
[26] D. P. Kingma, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 2015.