University College London

Department of Statistical Science

Scalable Inference for Generative Models using Quasi-Monte Carlo

Candidate number:
GMFG0

Supervisor:
Dr François-Xavier Briol

13 September 2020

Thesis in partial fulfilment of the requirements to obtain the degree:
MSc Data Science

12,547 words

Acknowledgements

Scientific research always builds on the shoulders of those who came before us and is dependent on their input, guidance and support for it to succeed. This holds even more true for me as this thesis was on the very edge of my prior knowledge and abilities. Therefore, I would like to thank my supervisor Dr François-Xavier Briol for his tireless and extraordinarily dedicated support. Your comments, feedback and guidance have elevated this thesis, my capabilities and my own confidence as a researcher.

Furthermore, my thanks go out to my corona bubble whose tenacity, resilience and optimism in unprecedented times have been an inspiration. Although no one would ask for a lockdown, your company has made the last six months an experience I will cherish and keep in my heart forever.

If the last year has taught me one thing, then it is that friendship cannot be measured in miles. We have been separated by borders and oceans, but you have never been closer to me and your continued support has meant the world to me.

Last but not least, there are those whose companionship we might take for granted at times. This goes out to my family and boyfriend: you have been by my side not only during this thesis, but stood by me at each turn and faced every challenge with me. Thank you.

Abstract

Modern computational statistics requires inference tools to scale well with model complexity and data size. In the specific case of inference for intractable generative models, the likelihood-free approach of minimum divergence estimators is applicable but typically relies on Monte Carlo (MC) methods to approximate the arising integrals. The convergence rate of this method is often considered to be too slow, motivating the use of the quasi-Monte Carlo (QMC) and randomised quasi-Monte Carlo (RQMC) methods which are constructed to improve upon the MC rate of convergence. This thesis explores the deployment QMC and RQMC methods to improve the scalability of a specific instance of minimum divergence estimators introduced by Briol et al. (2019) which is based on the maximum mean discrepancy (MMD). Empirical experiments on several classes of generative models show that the modified estimators perform at least as well as the MC-based one in terms of precision. For low-dimensional models and a large enough number of samples, faster convergence rates can be observed when using QMC and RQMC methods. Expanding this empirical analysis to a main competitor of the minimum MMD estimator, the Sinkhorn loss proposed by Genevay et al. (2018) is also found to profit from the use of QMC and RQMC point sets resulting in faster convergence rates for low dimensions. Thus, QMC methods are shown to improve the scalability of both considered minimum divergence estimators with respect to the data size for low-dimensional generative models.

Contents

List of Figures

Acronyms

ABC	approximate Bayesian computation
CLT	central limit theorem
GAN	generative adversarial networks
HK	Hardy-Krause
IID	identically and independently distributed
LLN	law of large numbers
MC	Monte Carlo
MCMC	Markov chain Monte Carlo
ML	maximum likelihood
MMD	maximum mean discrepancy
MSE	mean squared error
NSGD	natural stochastic gradient descent
QMC	quasi-Monte Carlo
RBF	radial basis function
RKHS	reproducing kernel Hilbert space
RMSE	root mean squared error
RQMC	randomised quasi-Monte Carlo
SGD	stochastic gradient descent
u.d. mod 1	uniformly distributed modulo 1

1 Introduction

When working with a parametric family of models, a natural goal is to infer estimates for the 'true' parameters. The classical approach to inference is maximum likelihood (ML) estimation which provides consistent and efficient parameter estimates (Norden 1973). However, the likelihood function can be intractable, i.e. expensive to evaluate or even impossible to define, if the model of interest is complex. Sometimes, this issue can be circumvented by approximating the likelihood using approaches such as pseudo likelihood (Besag 1974), profile likelihood (Murphy et al. 2000) or composite likelihood estimation (Varin et al. 2011). In the case of un-normalised models, the likelihood function is available up to a normalising constant and can be approximated using for instance Markov chain Monte Carlo (MCMC) algorithms (Møller et al. 2006).

For the family of intractable generative models, sometimes also referred to as implicit models or likelihood-free models, it is possible to obtain realisations given a specific parameter value, but the likelihood or an approximation thereof is not necessarily available (Mohamed et al. 2016). Thus, none of the aforementioned approaches is readily applicable. As generative models are widely used across the sciences including ecology (Ramstead et al. 2019), population genetics (Riesselman et al. 2018), or machine learning (Dziugaite et al. 2015; Y. Li et al. 2015), likelihood-free inference methods are required. One possible approach is the use of minimum divergence estimators which infer parameters by minimising some notion of distance between a 'true' and a candidate probability distribution (Basu et al. 1997).

Briol et al. (2019) proposed such a minimum divergence estimator for inference with generative models based on the maximum mean discrepancy (MMD). As opposed to other divergences, the MMD is easily computed even on large-scale datasets as it can be expressed in terms of the distance between kernel mean embeddings in a reproducing kernel Hilbert space (RKHS) (Gretton et al. 2012). This property makes the estimator applicable in complex, high-dimensional settings. Similar to other minimum divergence estimators, the minimum MMD estimator of Briol et al. (2019) relies on Monte Carlo (MC) methods for inference as pseudo-random points are utilised to generate samples from the generative model of interest. However, this approach fails to scale well with growing sample sizes as the convergence rate of MC is often considered to be too slow

for large-scale applications (Lemieux 2009, pp.11-12). Quasi-Monte Carlo (QMC) and randomised quasi-Monte Carlo (RQMC) methods remedy this drawback by relying on deterministic low-discrepancy points which are constructed to achieve faster convergence rates than MC. In this light, this thesis aims at empirically exploring possibilities to further improve the scalability of the minimum MMD estimator proposed by Briol et al. (2019) with respect to both model complexity and data size by deploying QMC and RQMC methods. In the attempt to leverage the higher efficiency of QMC and RQMC, numerical experiments are conducted in which pseudo-random point sets used for the simulation of samples from the generative model are replaced by deterministic low-discrepancy points.

Therefore, this thesis is organised as follows. Section 2 reviews RKHS and distances between probability distributions which are combined to introduce the general framework of minimum divergence estimators with a particular emphasis on inference for generative models. Then, the minimum MMD estimators of Briol et al. (2019) are covered in detail and a main competitor is briefly presented: estimators based on the minimum Sinkhorn divergence. Motivated by the arising integrals, section 3 gives an introduction of MC, QMC and RQMC methods used to approximate integration problems. In section 4, the novel application of QMC and RQMC methods to improve the scalability of the estimators in the minimum MMD framework of Briol et al. (2019) is introduced. The results of numerical experiments spanning three different generative models are presented in section 5 and their implications and limitations discussed in section 6. Section 7 concludes and suggests directions for future research.

2 Introduction to Minimum Divergence Estimators

Statistical models for which the likelihood function is intractable call for likelihood-free inference methods. Minimum divergence estimators fall into this category of approaches and can be based on various notions of distance. Here, the focus is on concepts describing distances between probability distributions. To be able to properly define those, the following section introduces reproducing kernel Hilbert spaces.

2.1 Reproducing Kernel Hilbert Spaces

A reproducing kernel Hilbert space (RKHS) is a class of functions. It is a special subclass of Hilbert spaces $\mathcal{H}$ which refer to vector spaces with the structure of an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ and an associated norm $\| \cdot \|_{\mathcal{H}} = \sqrt{\langle \cdot, \cdot \rangle_{\mathcal{H}}}$ (Kreyszig 1989, pp.128-129). The distinguishing characteristic of an RKHS is the reproducing property which it inherits from its associated reproducing kernel. This leads to two defining features of the RKHS $\mathcal{H}_k$: the kernel $k(\mathbf{x}, \mathbf{y}) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ has a feature map $k(\mathbf{x}, \cdot)$ of every point $\mathbf{x}$ in the metric space $\mathcal{X}$ which is in $\mathcal{H}_k$, and k satisfies the reproducing property which ensures that the evaluation of a function f at $\mathbf{x}$ can be expressed as an inner product in $\mathcal{H}_k$ between the function f and the feature map $k(\mathbf{x}, \cdot)$ (Muandet et al. 2017).

This reproducing property of an RKHS can convert many quantities of interest into tractable problems (Muandet et al. 2017). If this conversion is not directly possible, it may be feasible to utilise kernels to embed a given space into a different space in which a desired operation could be trivial to perform even if it is complex or intractable in the original space (Hofmann et al. 2008). The so-called 'kernel trick' implicitly performs this embedding by evaluating a kernel instead of computing an inner product in the original space, thereby avoiding a potentially infeasible direct computation (Schölkopf et al. 2002, p.34). This approach can be exploited to define distances between probability distributions by embedding the probability measures into an RKHS and then measuring the distance in the new space (Schölkopf et al. 2002, p.48).

Let $\mathcal{P}_k(\mathcal{X})$ be a set of Borel probability measures μ on the metric space $\mathcal{X}$. To transform the probability measure μ into an element of an RKHS, the kernel mean embedding $\Pi_k(\mu) = \int_{\mathcal{X}} k(\mathbf{x}, \cdot) \mathrm{d}\mu(\mathbf{x})$ of a symmetric and positive definite kernel is used (Muandet et al. 2017). Positive definite kernels satisfy the condition $\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \geq 0$ where $\forall n \in \mathbb{N}$, $\alpha_1 \ldots, \alpha_n \in \mathbb{R}^n$, and $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$ (Sriperumbudur et al. 2010), which is fulfilled when the kernel can be written as an inner product on $\mathcal{H}_k$ (Berlinet et al. 2004, p.7). If such a kernel is also characteristic, the kernel mean incorporates all information about the probability measure and can be used to define metrics on probability distributions (Muandet et al. 2017). A kernel is characteristic if the kernel mean embedding $\Pi_k(\mu)$ is injective, that is $\Pi_k(\mu)$ maps the kernel mean in $\mathcal{P}_k(\mathcal{X})$ to the distinct element $\int_{\mathcal{X}} k(\cdot, \mathbf{y}) \mu(\mathrm{d}\mathbf{y})$ in $\mathcal{H}_k$ (Sriperumbudur et al. 2011). As a sufficient condition

for a reproducing kernel to be characteristic, it has to be (i) measurable and bounded, i.e. $\sup_{\mathbf{x}} k(\mathbf{x}, \mathbf{x}) < \infty$, as well as (ii) integrally strictly positive definite, that is, for which $\int_{\mathcal{X}} \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{y}) \mu(\mathrm{d}\mathbf{x}) \mu(\mathrm{d}\mathbf{y}) = 0$ implies $\mu = 0$ for all $\mu \in \mathcal{P}_k$ (Sriperumbudur et al. 2010).

Examples of integrally strictly positive definite kernels include the Gaussian and the Laplacian kernel. The Gaussian kernel on $\mathbb{R}^d$ is defined as $k(\mathbf{x}, \mathbf{y}) = \exp(-\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{2\sigma^2})$ with lengthscale $\sigma > 0$ and $\|\cdot\|_2$ being the ℓ^2-norm, also known as the Euclidean distance (Muandet et al. 2017). These kernels belong to the class of radial basis function (RBF) kernels which have the form $k(\mathbf{x}, \mathbf{y}; l) = \gamma(\|\mathbf{x}-\mathbf{y}\|/l)$ with a radial function $\gamma : \mathbb{R} \to \mathbb{R}_{\geq 0}$ and lengthscale $l > 0$.

The following introduction of notions of distance will build upon the properties of RKHSs discussed above.

2.2 Distances between Probability Distributions

Kernel mean embeddings and RKHSs can be used to define distances between probability distributions and easily compute them. To explore this in more detail, we first introduce various concepts of distance.

Let $\mathcal{X}$ be a metric space and $\mathcal{P}(\mathcal{X})$ a set of Borel probability measures. The notion of *statistical divergence* refers to functions $D : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}_+$ satisfying for all $\mathbb{P}_1, \mathbb{P}_2 \in \mathcal{P}(\mathcal{X})$ the condition that $D(\mathbb{P}_1 \| \mathbb{P}_2) = 0$ if and only if $\mathbb{P}_1 = \mathbb{P}_2$ (Cichocki et al. 2010). Generally, divergences are hard or even impossible to calculate for large and potentially high-dimensional datasets (Park et al. 2016).

Metrics or pseudo-metrics can also be directly defined on probability measures. *Pseudo-probability metrics* refer to functions $d : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}_+$ satisfying the following axioms:

(i) $d(\mathbb{P}_1, \mathbb{P}_1) = 0$ (identity of indiscernibles)

(ii) $d(\mathbb{P}_1, \mathbb{P}_2) = d(\mathbb{P}_2, \mathbb{P}_1)$ (symmetry)

(iii) $d(\mathbb{P}_1, \mathbb{P}_2) \leq d(\mathbb{P}_1, \mathbb{P}_2) + d(\mathbb{P}_2, \mathbb{P}_3)$ (triangle inequality)

for all probability measures $\mathbb{P}_1, \mathbb{P}_2, \mathbb{P}_3 \in \mathcal{P}(\mathcal{X})$. *Probability metrics* are pseudo-probability metrics which satisfy $d(\mathbb{P}_1, \mathbb{P}_2) = 0$ if and only if $\mathbb{P}_1 = \mathbb{P}_2$ additionally to the conditions (i)-(iii) (Cohn 2013, p.86). Since statistical divergences are usually not symmetric and

do not satisfy the triangle inequality, probability metrics are divergences, but the reverse is not necessarily the case (Cichocki et al. 2010).

In the following, we will focus on two specific notions of distance, which are of special interest in the context of inference for generative models: the maximum mean discrepancy and the Wasserstein distance.

2.2.1 Maximum Mean Discrepancy

A common class of pseudo-probability metrics are *integral (pseudo-)probability metrics* (Müller 1997):

$$d_{\mathcal{F}}(\mathbb{P}_1 \| \mathbb{P}_2) = \sup_{f \in \mathcal{F}} \left| \int_{\mathcal{X}} f(\mathbf{x}) \mathbb{P}_1(\mathrm{d}\mathbf{x}) - \int_{\mathcal{X}} f(\mathbf{y}) \mathbb{P}_2(\mathrm{d}\mathbf{y}) \right| \tag{1}$$

where the function class $\mathcal{F}$ has to be chosen in the trade-off between being rich enough so that $d_{\mathcal{F}}(\mathbb{P}_1 \| \mathbb{P}_2) = 0$ if and only if $\mathbb{P}_1 = \mathbb{P}_2$, and being restrictive enough to allow for estimation (Muandet et al. 2017). The *maximum mean discrepancy (MMD)* is an integral probability metric for which the function class $\mathcal{F}$ corresponds to the unit ball in an RKHS, i.e. $\mathcal{F} := \{f \mid \|f\|_{\mathcal{H}_k} \leq 1\}$:

$$\mathrm{MMD}(\mathbb{P}_1 \| \mathbb{P}_2) = \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \left| \int_{\mathcal{X}} f(\mathbf{x}) \mathbb{P}_1(\mathrm{d}\mathbf{x}) - \int_{\mathcal{X}} f(\mathbf{y}) \mathbb{P}_2(\mathrm{d}\mathbf{y}) \right|.$$

Further, we can use the properties of an RKHS to obtain an alternative expression for the MMD by first embedding the probability distributions in an RKHS $\mathcal{H}_k$ and then using the distance in $\mathcal{H}_k$, i.e. the norm, to compare them: Borgwardt et al. (2006) show that the MMD can be expressed as the distance in $\mathcal{H}_k$ between mean embeddings such that $\mathrm{MMD}(\mathbb{P}_1 \| \mathbb{P}_2) = \|\Pi_k(\mathbb{P}_1) - \Pi_k(\mathbb{P}_2)\|_{\mathcal{H}_k}$ where $\|\cdot\|_{\mathcal{H}_k}$ is the norm associated with $\mathcal{H}_k$. The squared MMD, $\|\Pi_k(\mathbb{P}_1) - \Pi_k(\mathbb{P}_2)\|^2_{\mathcal{H}_k}$, can also be expressed in terms of the

associated kernel k using the reproducing property of $\mathcal{H}_k$ (Gretton et al. 2012):

$$\begin{aligned}\mathrm{MMD}^2(\mathbb{P}_1\|\mathbb{P}_2) &:= \left\| \int_{\mathcal{X}} k(\cdot,\mathbf{x})\mathbb{P}_1(\mathrm{d}\mathbf{x}) - \int_{\mathcal{X}} k(\cdot,\mathbf{x})\mathbb{P}_2(\mathrm{d}\mathbf{x}) \right\|^2_{\mathcal{H}_k} \\ &= \int_{\mathcal{X}}\int_{\mathcal{X}} k(\mathbf{x},\mathbf{x})\mathbb{P}_1(\mathrm{d}\mathbf{x})\mathbb{P}_1(\mathrm{d}\mathbf{x}) - 2\int_{\mathcal{X}}\int_{\mathcal{X}} k(\mathbf{x},\mathbf{y})\mathbb{P}_1(\mathrm{d}\mathbf{x})\mathbb{P}_2(\mathrm{d}\mathbf{y}) \\ &\quad + \int_{\mathcal{X}}\int_{\mathcal{X}} k(\mathbf{y},\mathbf{y})\mathbb{P}_2(\mathrm{d}\mathbf{y})\mathbb{P}_2(\mathrm{d}\mathbf{y}).\end{aligned} \tag{2}$$

This last expression for the squared MMD is simply a sum of expectations of the kernel k and can thus, dependent on the choice of k and the distributions $\mathbb{P}_1$ and $\mathbb{P}_2$, be calculated either in closed form or utilising numerical integration methods. For instance, a Gaussian kernel k combined with two Gaussian distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ would provide a closed form expression (Sriperumbudur et al. 2010). The expression in (2) also remedies the drawback of divergences being hard to calculate. Since the squared MMD can be easily computed using kernels, it is a suitable candidate for the use with complex models.

Building on the background of RKHS from the previous section, we can give conditions under which the MMD is a statistical divergence. Assuming that k is a bounded characteristic kernel, the kernel mean embedding in the corresponding RKHS norm $\|\cdot\|_{\mathcal{H}_k}$ satisfies $\mathrm{MMD}(\mathbb{P}_1\|\mathbb{P}_2) = \|\Pi_k(\mathbb{P}_1) - \Pi_k(\mathbb{P}_2)\|_{\mathcal{H}_k} = 0$ for $\mathbb{P}_1, \mathbb{P}_2 \in \mathcal{P}_k(\mathcal{X})$ if and only if $\mathbb{P}_1 = \mathbb{P}_2$. This condition corresponds to the additional axiom presented in section 2.2, which has to hold for a pseudo-probability metric to be a probability metric. Hence, the norm of the RKHS, the MMD, is such a probability metric on $\mathcal{P}_k$ and therefore also a statistical divergence (Muandet et al. 2017). This condition is assumed to hold throughout the following sections.

The MMD can be used in various applications and is especially popular in testing. Naturally, the MMD is suitable for testing the equality of two samples and has the advantage of being distribution-free, i.e. no assumption on the parametric form of the distribution is required (Muandet et al. 2017). Borgwardt et al. (2006), for instance, develop such an MMD-based two-sample test to address the problem of data integration in bioinformatics. Goodness-of-fit tests, which test whether a particular sample is drawn from a given distribution, can also be turned into a two-sample problem and then analysed using an MMD-based test statistic (Muandet et al. 2017). The problem of distribution comparison might even be interpreted as a binary classification problem: the

loss function of a classifier such as a neural network or support vector machine can be defined in terms of MMD (Fukumizu et al. 2009). Further, the MMD enjoys popularity in the context of approximate Bayesian computation (ABC) which is a Bayesian inference paradigm for cases in which the likelihood is intractable but generating data for given parameters is simple. Here, the MMD serves as a measure of similarity between the pseudo data and the observed data defining the approximate posterior (Park et al. 2016). Another application of MMD, which is in the context of statistical inference for generative models, will be explored further in the following sections.

2.2.2 Wasserstein Distance

A different metric useful for parameter estimation is the *Wasserstein distance*, also called earth mover's distance, which is based on the theory of optimal transport. The theory of optimal mass transport has roots in civil engineering (Monge 1781) and economics (Kantorovitch 1958). The original mass transportation problem of finding an optimal way to move a pile of soil from one site to another by minimising the required transportation cost was formulated by Monge (1781). Kantorovitch (1958) proposed a relaxation of this nonlinear problem, thereby converting it into a linear programming problem (Villani 2003, p.3). Let $\mathbb{P}_1$ and $\mathbb{P}_2$ be Borel probability measures on the metric space $\mathcal{X}$ and let $c(\mathbf{x}, \mathbf{y})$ denote the transportation costs of moving one unit of mass from $\mathbf{x}$ to $\mathbf{y}$ for $\mathbf{x} \neq \mathbf{y} \in \mathcal{X}$. Then, the optimal mass transport problem aims to find the transport map minimising the total transportation costs. Following the formulation of Kantorovitch (1958), the optimal transport distance is the p-Wasserstein distance:

$$W_p(\mathbb{P}_1, \mathbb{P}_2) = \left(\inf_{J \in \mathcal{J}(\mathbb{P}_1, \mathbb{P}_2)} \int_{\mathcal{X} \times \mathcal{X}} c(\mathbf{x}, \mathbf{y}) \mathrm{d}J(\mathbf{x}, \mathbf{y}) \right)^{1/p}$$

where $\mathcal{J}(\mathbb{P}_1, \mathbb{P}_2)$ denotes the set of all joint probability measures J on $\mathcal{X} \times \mathcal{X}$ with marginals $\mathbb{P}_1$ and $\mathbb{P}_2$. Commonly, the transportation costs $c(\mathbf{x}, \mathbf{y})$ are defined in terms of distance, i.e. the ℓ^p-norm $\|\cdot\|_p$. For example, the 1-Wasserstein distance has $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$. This 1-Wasserstein distance is also an instance of an integral probability metric of equation (1), where $\mathcal{F}$ is obtained from the unit ball of 1-Lipschitz functions such that $\mathcal{F} := \{f | \sup_{\mathbf{x} \neq \mathbf{y} \in \mathcal{X}} |f(\mathbf{x}) - f(\mathbf{y})| / \|\mathbf{x} - \mathbf{y}\| \leq 1\}$ (Muandet et al. 2017).

Wasserstein distances have widely been explored in computer vision and machine learning to compare histograms in order to pool local features for a variety of purposes: in image retrieval applications, the Wasserstein distance is used to measure the dissimilarity between image signatures (Rubner et al. 2000; P. Li et al. 2013), whereas in image registration, it is employed to find a common geometric reference between images (Haker et al. 2004). Arjovsky et al. (2017) use Wasserstein distances to train generative adversarial networks (GANs), while El Moselhy et al. (2012) and Ramdas et al. (2017) utilise them in the context of Bayesian inference and two-sample testing respectively.

Hence, it is clear that Wasserstein distances represent a powerful tool for the comparison of probability distributions. However, they suffer from two main disadvantages: from a computational point of view, Wasserstein distances are expensive to compute since they require solving a linear program for which the worst case time complexity is exponential (Rubner et al. 2000). More efficient algorithms still have a time complexity of $\mathcal{O}(n^3 \log n)$ (Pele et al. 2009). Considering the statistical view point, they suffer from the curse of dimensionality due to their slow convergence (Genevay et al. 2019). A d-dimensional empirical measure has been shown to converge to its true measure in the 1-Wasserstein distance at $n^{-1/d}$ implying that the convergence rate slows down considerably for high-dimensional data (Weed et al. 2019).

Recent works have exploited the use of entropic regularisation to decrease the computational cost. Depending on the strength of the regularisation ϵ, the so-called *Sinkhorn divergence* interpolates between unregularised Wasserstein distance ($\epsilon = 0$) and MMD ($\epsilon = +\infty$) (Genevay et al. 2019). Genevay et al. (2018) introduce the regularised optimal transport problem using the definition

$$J_\epsilon := \inf_{J \in \mathcal{J}(\mathbf{x},\mathbf{y})} \int_{\mathcal{X}\times\mathcal{X}} c(\mathbf{x},\mathbf{y}) \mathrm{d}J(\mathbf{x},\mathbf{y}) + \epsilon \int_{\mathcal{X}\times\mathcal{X}} \log\left(\frac{J(\mathbf{x},\mathbf{y})}{\mathrm{d}\mathbb{P}_1(\mathbf{x})\mathrm{d}\mathbb{P}_2(\mathbf{y})}\right) \mathrm{d}J(\mathbf{x},\mathbf{y}).$$

The associated regularised Wasserstein distance can then be expressed as $W_{c,\epsilon}(\mathbb{P}_1,\mathbb{P}_2) = \int_{\mathcal{X}\times\mathcal{X}} c(\mathbf{x},\mathbf{y})\mathrm{d}J_\epsilon(\mathbf{x},\mathbf{y})$ where J_ϵ is the optimal coupling for the regularised optimal transport problem defined above. To calculate the Sinkhorn divergence very quickly, the Sinkhorn algorithm can be used which benefits from a linear convergence rate (Peyré et al. 2019). Consider Cuturi (2013) for a detailed discussion of the matter.

2.3 Minimum Divergence Estimators

Having introduced various distances between probability distributions, we now aim at deploying these to tackle the problem of statistical inference. A common approach is to use minimum divergence estimators which intend to find the parameters minimising some notion of divergence between the true empirical distribution and a candidate distribution (Basu et al. 1998). Formally, we are interested in inference with a parametric family $\mathcal{P}_\Theta(\mathcal{X}) = \{\mathbb{P}_\theta \in \mathcal{P}(\mathcal{X}) : \theta \in \Theta\}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{P}(\mathcal{X})$ denotes the set of Borel probability measures on this metric space (Basu et al. 2011, p.27).

The class of minimum divergence estimators minimises some loss function based on a divergence between a parametric model $\mathbb{P}_\theta$ and an empirical probability measure $\mathbb{Q}^m = \frac{1}{m}\sum_{j=1}^m \delta_{\mathbf{y}_j}$ where $\delta_{\mathbf{y}_j}$ is the Dirac measure and $\{\mathbf{y}_j\}_{j=1}^m \overset{\text{IID}}{\sim} \mathbb{Q}$. A Dirac measure at $a \subset \mathcal{X}$ is defined as $\delta_a(A) = 1$ if $a \in A$ and 0 otherwise (Tao 2011, p.90). Thus, the estimators are obtained as the solution of the following optimisation problem (Basu et al. 2011, p.27):

$$\hat{\theta}_m = \arg\min_{\theta\in\Theta} D(\mathbb{P}_\theta \| \mathbb{Q}^m).$$

In this context, two scenarios are of interest: the M-closed and the M-open setting. In the M-closed setting, we assume that $\mathbb{Q} \in \mathcal{P}_\Theta(\mathcal{X})$, i.e. the considered probability measure $\mathbb{Q}$ is in the parametric family $\mathcal{P}_\Theta(\mathcal{X})$. In the M-open setting, $\mathbb{Q}$ does not necessarily need to be an instance of $\mathcal{P}_\Theta(\mathcal{X})$, but can be any probability measure in $\mathcal{P}(\mathcal{X})$ (Briol et al. 2019). Even though the M-closed setting is more restrictive, it can be analysed more easily from a theoretical point of view. The M-open setting is closer to real-world applications, but implies the need to examine the robustness of the estimator (Bernardo et al. 1994, pp. 384-385).

Minimum divergence estimators have a wide spectrum of applications encompassing areas such as machine learning (e.g. Dziugaite et al. 2015; Y. Li et al. 2015) and Bayesian computation (e.g. Park et al. 2016; Ratmann et al. 2007). Also, they are closely connected to the concept of scoring rules (Dawid 2007; Gneiting et al. 2007). Such a scoring rule quantifies the accuracy of a model and if it is strictly proper, that is the score is minimised when the distribution of generated and observed data is equivalent, it induces a divergence which can be used in a loss function to obtain minimum divergence

estimators (Steinwart et al. 2019).

In the following, a specific application of minimum divergence estimators will be considered: the application to the class of generative models. For generative models, the explicit likelihood is not available, but we assume that for any value of the parameter vector $\theta \in \Theta$ it is possible to generate identically and independently distributed (IID) data. A generative model is considered a probability measure for which we can obtain IID samples $\{\mathbf{x}_i\}_{i=1}^n$, where $x \in \mathcal{X}$, from the corresponding probability measure $\mathbb{P}_\theta$ for any given parameter $\theta \in \Theta$. To generate samples, IID realisations $\{\mathbf{u}_i\}_{i=1}^n$, where $u \in \mathcal{U}$ and $\mathcal{U} \subseteq \mathbb{R}^d$, are attained from the probability measure $\mathbb{U}$ and then mapped through a differentiable function $G_\theta : \mathcal{U} \to \mathcal{X}$, the generator, to produce the $\mathbb{P}_\theta$ distributed independent samples $\mathbf{x}_i = G_\theta(\mathbf{u}_i)$ for $i = 1, \ldots, n$ (Genevay et al. 2018).

The application of minimum distance estimators for generative models is based on the idea of simulating data using varying parameters to find the specific parameter generating the dataset that best matches the true observations in terms of its distribution. The available methods differ in the choice of the used divergence. Examples include the Wasserstein distance (Bassetti et al. 2006) and its approximation using the Sinkhorn relaxation (Genevay et al. 2018), or the MMD (Briol et al. 2019). In practice, minimum divergence estimators are often costly to implement as applying distances directly on a dataset is difficult when the dataset is large and possibly high-dimensional (Park et al. 2016). In the following, we therefore focus in detail on one specific distance, the MMD, which can be expressed using kernel methods and is thus easily computed.

2.3.1 Minimum MMD Estimators

Briol et al. (2019) propose a minimum divergence estimator framework based on the squared MMD between an unknown data generation distribution $\mathbb{Q}$ and an instance of the parametric family of model distributions $\mathbb{P}_\theta \in \mathcal{P}_\Theta(\mathcal{X})$. A minimum MMD estimator requires an empirical estimate of the squared MMD which can be based on U-statistics. Such a U-statistic is obtained from a symmetric average of the kernel k over all observations and gives an unbiased estimate of the approximated function (Serfling 1980, pp.171-173). Using this approach, Briol et al. (2019) consider the following minimum

MMD estimator:

$$\hat{\theta}_m = \arg\min_{\theta \in \Theta} \mathrm{MMD}_U^2(\mathbb{P}_\theta \| \mathbb{Q}^m) \tag{3}$$

where $\mathbb{Q}^m(\mathrm{d}\mathbf{y}) = \frac{1}{m}\sum_{i=1}^m \delta_{\mathbf{y}_i}$ with Dirac measure $\delta_{\mathbf{y}_i}$ and $\{\mathbf{y}_i\}_{i=1}^m \overset{\text{IID}}{\sim} \mathbb{Q}$. This implies that the approximated squared MMD has the following form:

$$\begin{aligned}\mathrm{MMD}_U^2(\mathbb{P}_\theta \| \mathbb{Q}^m) = \int_{\mathcal{X}} \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{x}) \mathbb{P}_\theta(\mathrm{d}\mathbf{x}) \mathbb{P}_\theta(\mathrm{d}\mathbf{x}) - \frac{2}{m} \sum_{j=1}^m \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{y}_j) \mathbb{P}_\theta(\mathrm{d}\mathbf{x}) \\ + \frac{1}{m(m-1)} \sum_{j \neq j'} k(\mathbf{y}_j, \mathbf{y}_{j'})\end{aligned}.$$

In practice, the generator of the model and the corresponding gradient may be expensive to evaluate, so that the number of simulated samples from $\mathbb{P}_\theta$ cannot be chosen arbitrarily large in order to avoid intractability. Thus, a possibly limited number of simulated samples n in practice motivates the study of a finite sample estimate of the squared MMD based on a second U-statistic approximation and the corresponding minimum MMD estimator:

$$\begin{aligned}\hat{\theta}_{n,m} &= \arg\min_{\theta \in \Theta} \mathrm{MMD}_{U,U}^2(\mathbb{P}_\theta^n \| \mathbb{Q}^m) \\ \mathrm{MMD}_{U,U}^2(\mathbb{P}_\theta^n \| \mathbb{Q}^m) &= \frac{1}{n(n-1)} \sum_{i \neq i'} k(\mathbf{x}_i, \mathbf{x}_{i'}) - \frac{2}{mn} \sum_{j=1}^m \sum_{i=1}^n k(\mathbf{x}_i, \mathbf{y}_j) \\ &\quad + \frac{1}{m(m-1)} \sum_{j \neq j'} k(\mathbf{y}_j, \mathbf{y}_{j'})\end{aligned}$$

where $\mathbb{P}_\theta^n = \frac{1}{n}\sum_{i=1}^n \delta(\mathbf{x}_i)$ and $\{\mathbf{x}_i\}_{i=1}^n \overset{\text{IID}}{\sim} \mathbb{P}_\theta$. The U-statistic approximation represents an unbiased estimator of the squared MMD, i.e. $\mathbb{E}[\mathrm{MMD}_{U,U}^2(\mathbb{P}_\theta^n \| \mathbb{Q}^m)] = \mathrm{MMD}^2(\mathbb{P}_\theta \| \mathbb{Q})$. This general framework of minimum MMD estimators is applied in various areas ranging from kernel scoring rules (Gneiting et al. 2007) to MMD generative adversarial networks (Y. Li et al. 2015; Dziugaite et al. 2015), and, in the Bayesian context, kernel ABC (Fukumizu et al. 2013; Park et al. 2016).

Numerical Optimisation

Generally, the optimisation problem in equation (3) will be non-convex and the minimiser $\hat{\theta}_m$ impossible to retrieve analytically. Thus, the gradient of the squared MMD,

$\nabla_\theta \text{MMD}^2(\mathbb{P}_\theta \| \mathbb{Q}^m)$ where $\nabla_\theta = (\partial_{\theta_1}, \ldots, \partial_{\theta_p})$, is required in order to be able to take a numerical optimisation approach. Assuming that the generator G_θ can be differentiated with respect to θ with a $\mathbb{U}$-integrable Jacobian matrix $\nabla_\theta G_\theta$, the gradient of the squared MMD can be expressed as

$$\begin{aligned}\nabla_\theta \text{MMD}^2(\mathbb{P}_\theta \| \mathbb{Q}^m) = 2 \int_{\mathcal{U}} \int_{\mathcal{U}} \nabla_1 k(G_\theta(\mathbf{u}), G_\theta(\mathbf{v})) \nabla_\theta G_\theta(\mathbf{u}) \mathbb{U}(\mathrm{d}\mathbf{u}) \mathbb{U}(\mathrm{d}\mathbf{v}) \\ - \frac{2}{m} \sum_{j=1}^{m} \int_{\mathcal{U}} \nabla_1 k(G_\theta(\mathbf{u}), \mathbf{y}_j) \nabla_\theta G_\theta(\mathbf{u}) \mathbb{U}(\mathrm{d}\mathbf{u})\end{aligned}$$

where $\nabla_1 k$ denotes the partial derivative with respect to the first argument. In practice, the integral terms cannot be determined analytically. However, it is again possible to approximate the gradient using U-statistics:

$$\begin{aligned}\hat{J}(\mathbb{Q}^m) = \frac{2}{n(n-1)} \sum_{i \neq i'} \nabla_\theta G_\theta(\mathbf{u}_i) \nabla_1 k(G_\theta(\mathbf{u}_i), G_\theta(\mathbf{u}_{i'})) \\ - \frac{2}{nm} \sum_{j=1}^{m} \sum_{i=1}^{n} G_\theta(\mathbf{u}_i) \nabla_1 k(G_\theta(\mathbf{u}_i), \mathbf{y}_j)\end{aligned}$$

where $\{\mathbf{u}_i\}_{i=1}^n \overset{\text{IID}}{\sim} \mathbb{U}$. This approximation is an unbiased estimator such that $\mathbb{E}[\hat{J}(\mathbb{Q}^m)] = \nabla_\theta \text{MMD}^2(\mathbb{P}_\theta \| \mathbb{Q}^m)$ where the expectation is taken over the IID realisations $\mathbf{u}_i$.

Given the gradient approximation $\hat{J}(\mathbb{Q}^m)$, gradient-based optimisation algorithms can be considered to find the minimum MMD estimate $\hat{\theta}$. A commonly used method is stochastic gradient descent (SGD). After initialising at $\hat{\theta}^{(0)} \in \Theta$, the algorithm iterates over the following steps:

1. sample $\{\mathbf{u}_i\}_{i=1}^n \overset{\text{IID}}{\sim} \mathbb{U}$ and compute $\mathbf{x}_i = G_{\hat{\theta}^{(t-1)}}(\mathbf{u}_i)$ for $i = 1, \ldots, n$
2. compute $\hat{\theta}^{(t)} = \hat{\theta}^{(t-1)} - \eta_k \hat{J}_{\hat{\theta}^{(t-1)}}(\mathbb{Q}^m)$

where the step size $(\eta_t)_{t \in \mathbb{N}}$ is chosen to ensure convergence to a local minimum. For $\mathcal{X} \subseteq \mathbb{R}^d$ and $\Theta \subseteq \mathbb{R}^p$, the computational cost per iteration of the SGD algorithm is $O((n^2 + nm)dp)$ implying that the costs increase linearly in the number of true data points m, but quadratically in number of simulated data points n.

Although SGD is supposed to iteratively decrease the value of the objective function, it does not move into the optimal direction in the considered RKHS, since we optimise

within a statistical manifold, in which notions of distance may differ from those in common Euclidean spaces (Amari 2016, p.4). SGD can still be used to find the local optimum of the loss function, but the algorithm will need a large number of iterations. In such a statistical manifold, each point is a Borel probability measure $\mathbb{P} \in \mathcal{P}(\mathcal{X})$. Manifolds commonly correspond to parametric families $\mathcal{P}_\Theta(\mathcal{X}) \subset \mathcal{P}(\mathcal{X})$ representing classes of probability measures $\mathbb{P}_\theta$ with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) \in \Theta$ giving a possible coordinate system within the statistical manifold (Amari 2016, p.3). A notion of distance between the coordinates is introduced by the choice of divergence. This divergence induces a symmetric metric tensor $g(\theta)$ consisting of a positive semi-definite matrix $(g_{ij}(\theta))$. Given that $g_{ij}(\theta)$ is positive definite at every $\theta \in \Theta$ for $i, j \in \{1, \ldots, p\}$, the function $g(\theta)$ maps the parameter θ into the local coordinate system $g_{ij}(\theta)$ (Amari 2016, p.10).

Here, the statistical manifold is induced by the MMD over the parametric family of probability measures $\mathcal{P}_\Theta(\mathcal{X})$ where the parameter space Θ is assumed to be a subset of $\mathbb{R}^p$ for $p \in \mathbb{N}$. Briol et al. (2019) show that the metric tensor induced by MMD corresponds to the information metric associated with the squared MMD divergence and can be expressed as

$$g(\theta) = \int_{\mathcal{U}} \int_{\mathcal{U}} \nabla_\theta G_\theta(\mathbf{u})^\top \nabla_2 \nabla_1 k\Big(G_\theta(\mathbf{u}), G_\theta(\mathbf{v})\Big) \nabla_\theta G_\theta(\mathbf{v}) \mathbb{U}(\mathrm{d}\mathbf{u}) \mathbb{U}(\mathrm{d}\mathbf{v}) \tag{4}$$

where $\nabla_2 \nabla_1 k(\mathbf{x}, \mathbf{y}) = \{\partial_{\mathbf{y}_j} \partial_{\mathbf{x}_i} k(\mathbf{x}, \mathbf{y})\}_{i,j=1,\ldots,d}$. In case of MMD, $g(\theta)$ cannot be evaluated exactly as the expression in (4) involves intractable integrals against $\mathbb{U}$. However, the information metric can be approximated using a U-statistic:

$$g_U(\theta) = \frac{1}{n(n-1)} \sum_{i \neq j} \nabla_\theta G_\theta(\mathbf{u}_i)^\top \nabla_2 \nabla_1 k\Big(G_\theta(\mathbf{u}_i), G_\theta(\mathbf{u}_j)\Big) \nabla_\theta G_\theta(\mathbf{u}_j)$$

where $\{\mathbf{u}_i\}_{i=1}^n \overset{\text{IID}}{\sim} \mathbb{U}$.

Briol et al. (2019) introduce a natural stochastic gradient descent (NSGD) algorithm which can leverage the geometry of the MMD statistical manifold to obtain more efficient updates. The optimal gradient descent direction for a loss function L in this manifold, i.e. the gradient, corresponds to the vector field $\nabla^g L$ which can be mapped into the local coordinate system using the inverse of $g(\theta)$: $\nabla^d L(\theta) = g^{-1}(\theta) \nabla_\theta L(\theta)$ (Amari 1998).

Given the approximation of the metric tensor $g_U(\theta)$, the objective function can now be optimised after initialising at $\hat{\theta}^{(0)} \in \Theta$ through iterating over the following descent steps:

1. sample $\{\mathbf{u}_i\}_{i=1}^n \overset{\text{IID}}{\sim} \mathbb{U}$ and compute $\mathbf{x}_i = G_{\hat{\theta}^{(t-1)}}(\mathbf{u}_i)$ for $i = 1, \ldots, n$
2. compute $\hat{\theta}^{(t)} = \hat{\theta}^{(t-1)} - \eta_k \, g_U\big(\hat{\theta}^{(t-1)}\big)^{-1} \hat{J}_{\hat{\theta}^{(t-1)}}(\mathbb{Q}^m)$.

This optimisation algorithm can be interpreted as an SGD procedure using pre-conditioning. Significant computational gains can be obtained despite the additional computational costs of $O((n^2 + nm)p^2 d + p^3)$ caused by the inversion of the approximate metric tensor.

Statistical Properties

Briol et al. (2019) restrict the analysis of the statistical properties of the estimators $\hat{\theta}_m$ and $\hat{\theta}_{m,n}$ to settings in which $\mathcal{X} \subset \mathbb{R}^d$ and $\Theta \subset \mathbb{R}^p$ for $d, p \in N$, i.e. to the M-closed case. The effects on efficiency, consistency and robustness will be presented in the following.

Since the optimisation goal is to minimise an empirical estimate of the squared MMD, either $\text{MMD}_U^2(\mathbb{P}_\theta \| \mathbb{Q}^m)$ or $\text{MMD}_{U,U}^2(\mathbb{P}_\theta^n \| \mathbb{Q}^m)$, the resulting estimate is subject to Monte Carlo error. This implies that the optimal parameters found to minimise the empirical estimate of the squared MMD, do not necessarily generalise to minimising the exact population squared MMD, $\text{MMD}^2(\mathbb{P}_\theta \| \mathbb{Q})$ (Dziugaite et al. 2015). Briol et al. (2019) show the generalisation error measured in terms of MMD to be bounded for both estimators $\hat{\theta}_m$ and $\hat{\theta}_{m,n}$ under mild conditions and that this bound is only dependent on n and m. Additionally assuming that the kernel k is bounded, the following generalisation error bounds hold for $\hat{\theta}_m$ and $\hat{\theta}_{m,n}$ respectively at least with probability $1 - \delta$ (Briol et al. 2019, Theorem 1):

$$\text{MMD}(\mathbb{P}_{\hat{\theta}_m} \| \mathbb{Q}) \leq \inf_{\theta \in \Theta} \text{MMD}(\mathbb{P}_\theta \| \mathbb{Q}) + 2\sqrt{\frac{2}{m} \sup_{x \in \mathcal{X}} k(\mathbf{x}, \mathbf{x})} \left(2 + \sqrt{\log\left(\frac{1}{\delta}\right)}\right),$$

$$\begin{aligned}\text{MMD}(\mathbb{P}_{\hat{\theta}_{m,n}} \| \mathbb{Q}) \leq \inf_{\theta \in \Theta} \text{MMD}(\mathbb{P}_\theta \| \mathbb{Q}) \\ + 2\left(\sqrt{\frac{2}{n}} + \sqrt{\frac{2}{m}}\right) \sqrt{\sup_{x \in \mathcal{X}} k(\mathbf{x}, \mathbf{x})} \left(2 + \sqrt{\log\left(\frac{1}{\delta}\right)}\right).\end{aligned}$$

It appears that the generalisation errors decrease independently of the dimensions p and

d as $n^{-1/2}$ and $m^{-1/2}$. Further, assuming that the kernel is bounded, these rates do not depend on the its properties. In fact, given a translation invariant kernel, $k(\mathbf{x}, \mathbf{x})$ is simply the maximum value of the kernel. The $O(n^{-1/2})$ rate of convergence in the generalisation errors is driven by how quickly an empirical measure $\mathbb{P}^n$, which is based on a $\mathbb{P}$-distributed IID sample of n data points, converges to the probability measure $\mathbb{P}$ on $\mathcal{X} \subseteq \mathbb{R}^d$ in terms of their MMD. The relevant probabilistic bound is given by the following concentration inequality (Briol et al. 2019, Lemma 1):

$$\text{MMD}(\mathbb{P}\|\mathbb{P}^n) \leq \sqrt{\frac{2}{n} \sup_{x \in \mathcal{X}} k(\mathbf{x}, \mathbf{x})} \left(1 + \sqrt{\log\left(\frac{1}{\delta}\right)}\right). \tag{5}$$

Again, this bound decreases as $n^{-1/2}$.

Under the previously stated assumptions and the existence of a unique minimiser $\theta^* \in \Theta$ such that $\text{MMD}(\mathbb{P}_{\theta^*}\|\mathbb{Q}) = \inf_{\theta \in \Theta} \text{MMD}(\mathbb{P}_\theta\|\mathbb{Q})$, Briol et al. (2019, Proposition 1) show that both considered estimators are consistent, implying $\lim_{m\to\infty} \hat{\theta}_m = \theta^*$ and $\lim_{m,n\to\infty} \hat{\theta}_{m,n} = \theta^*$ almost surely. This result provides conditions under which the bounds identified above convert into convergence of the estimators. However, the estimators are not considered efficient, and, thus, are not expected to outperform a ML approach in the M-closed setting. Further, it can be noted that the efficiency of $\hat{\theta}_m$ and $\hat{\theta}_{m,n}$ strongly depends on the choice of the kernel and especially its corresponding lengthscale l.

So far, the discussion of the estimator properties has been limited to the M-closed case. In practice, this assumption does not necessarily hold, as it is common to encounter corrupted data. For generative models, an estimator which is non-sensitive to small deviations from the assumed true distribution is desirable (Huber et al. 2011, p.1). Therefore, Briol et al. (2019) study the robustness of the estimators in the M-open setting. Both estimators $\hat{\theta}_m$ and $\hat{\theta}_{m,n}$ are found to be qualitatively robust in the sense of Hampel (1971) and, under additional assumptions, also bias-robust (see e.g. Huber et al. 2011, pp.13-15).

2.3.2 Minimum Sinkhorn Divergence Estimators

A main competitor for minimum MMD estimators are minimum divergence estimators based on the Wasserstein distance and its Sinkhorn relaxation as introduced in section

2.2.2. Genevay et al. (2018), for instance, consider the following estimator

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \overline{W}_{c,\epsilon}(\mathbb{P}_\theta^n, \mathbb{Q}^m)$$

where $\mathbb{Q}^m = \frac{1}{m}\sum_{i=1}^m \delta_{\mathbf{y}_i}$ and $\{\mathbf{y}_i\}_{i=1}^m \overset{\text{IID}}{\sim} \mathbb{Q}$, as well as $\mathbb{P}_\theta^n = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}$ and $\{\mathbf{x}_i\}_{i=1}^n \overset{\text{IID}}{\sim} \mathbb{P}_\theta$. This estimator is based on the *Sinkhorn loss* between the two measures $\mathbb{P}_1$ and $\mathbb{P}_2$ which is defined as (Genevay et al. 2018, Theorem 1):

$$\overline{W}_{c,\epsilon}(\mathbb{P}_1, \mathbb{P}_2) = 2W_{c,\epsilon}(\mathbb{P}_1, \mathbb{P}_2) - W_{c,\epsilon}(\mathbb{P}_1, \mathbb{P}_1) - W_{c,\epsilon}(\mathbb{P}_2, \mathbb{P}_2).$$

Depending on the regularisation parameter ϵ, this Sinkhorn loss interpolates between the double Wasserstein distance of measures $\mathbb{P}_1$ and $\mathbb{P}_2$, i.e. $\overline{W}_{c,\epsilon}(\mathbb{P}_1, \mathbb{P}_2) \to 2W_c(\mathbb{P}_1, \mathbb{P}_2)$ as $\epsilon \to 0$, and their MMD with the kernel corresponding to the cost from the optimal transport problem, that is $\overline{W}_{c,\epsilon}(\mathbb{P}_1, \mathbb{P}_2) \to \text{MMD}_{-c}(\mathbb{P}_1, \mathbb{P}_2)$ as $\epsilon \to +\infty$. Due to this property, the estimators based on this Sinkhorn loss can leverage the sample complexity of $O(n^{-1/2})$ of MMD losses, thereby improving upon the sample complexity of $O(n^{-1/d})$ for unregularised Wasserstein losses (Genevay et al. 2018). The cost function c can be chosen freely and is thus not restricted to positive definite kernels as it is the case for minimum MMD estimators. The regularisation parameter ϵ can be tuned to optimally combine the geometry of Wasserstein distances and the high-dimensional rigidity of MMD-based losses. Then, the estimation procedure follows a combination of gradient descent and a loss approximation using the Sinkhorn algorithm. For details see Genevay et al. (2018).

3 Introduction to Monte Carlo and Quasi-Monte Carlo

As we have already seen in the previous section, integrals are ubiquitous in various applications in both Bayesian and frequentist statistics. Bayesian inference for instance relies on the marginal likelihood, an integral, to exactly evaluate the posterior density on the parameters of interest (Gelman et al. 2014, p.261). The computation of integrals is also often required for classical ML estimation in order to maximise the expected log-likelihood under a certain data-generating process (Heiss et al. 2008). In many cases,

these integrals cannot be computed in closed form, so that we need to resort to numerical integration methods. Thus, the numerical integration of complex and computationally expensive functions can be considered as one of the major challenges in computational statistics and is, in the specific case of this thesis, required to obtain the minimum divergence estimators of interest. The following sections will focus on stochastic methods for numerical integration, which approximate possibly high-dimensional integrals more effectively than deterministic approaches (Kuo et al. 2005). The formal presentation of Monte Carlo and quasi-Monte Carlo integration requires some background in measure and probability theory for the definitions of Riemann and Lebesgue integrals. An introduction including those definitions can be found in appendix A.

3.1 Monte Carlo Methods

Considering a metric space $\mathcal{X}$ and a set of Borel probability measures $\mathcal{P}(\mathcal{X})$, the formal goal for MC integration is to approximate the Lebesgue integral of some integrable function $f : \mathcal{X} \to \mathbb{R}$ (Niederreiter 1992, pp.3-4):

$$I_d(f) = \int_{\mathcal{X}} f(\mathbf{x})\mathbb{P}(\mathrm{d}\mathbf{x}) \tag{6}$$

where $\mathbb{P} \in \mathcal{P}(\mathcal{X})$. The MC method approximates this integral by computing the average of the considered function evaluated on a set of pseudo-random points. Thus, it is based on the idea of estimating the population expectation by the sample mean. The required evaluation points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are IID samples from the probability measure $\mathbb{P}$ (Niederreiter 1992, p.5). For the approximation, the MC method uses an equal-weight cubature rule of the form

$$Q_{n,d}(f) = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i) \tag{7}$$

where $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,d})$. Thus, $Q_{n,d}(f)$ can also be treated as a random variable with mean and variance (Lemieux 2009, p.10). The mean of $Q_{n,d}(f)$ can be proven to be

$$\mathbb{E}\left[Q_{n,d}(f)\right] = I_d(f)$$

implying that the MC method is unbiased, and its variance is given by

$$\mathrm{Var}\left[Q_{n,d}(f)\right] = \mathbb{E}\left[\left|Q_{n,d}(f) - I_d(f)\right|^2\right] = \frac{\sigma^2(f)}{n}$$

where $\sigma^2(f) := I_d(f^2) - (I_d(f))^2$ is the variance of f.

To examine how accurate standard MC is, two fundamental results of probability theory can be consulted: the law of large numbers (LLN) and the central limit theorem (CLT). Assuming that $I_d(f)$ exists, the strong LLN ensures that

$$\mathbb{P}\left(\lim_{n\to\infty} |Q_{n,d}(f) - I_d(f)| = 0\right) = 1.$$

Therefore, under broad assumptions, the sample mean $Q_{n,s}(f)$ will get arbitrarily close to its expectation $I_d(f)$ as $n \to \infty$ (Niederreiter 1992, p.4). This result, however, gives no information about the rate at which this happens. Assuming that the function f is square-integrable with respect to $\mathbb{P}$, i.e. $\int_{\mathcal{X}} |f(x)|^2 \mathbb{P}(\mathrm{d}x) < \infty$, to guarantee a finite variance of f, the CLT provides this rate of convergence. The standard deviation of the absolute error $|I_d(f) - Q_{n,d}(f)|$ can be obtained from

$$\sqrt{\mathbb{E}\left[\left|I_d(f) - Q_{n,d}(f)\right|^2\right]} = \frac{\sigma(f)}{\sqrt{n}} \tag{8}$$

and is often called the root mean squared error (RMSE). The CLT states that

$$\lim_{n\to\infty} \mathbb{P}\left(|I_d(f) - Q_{n,d}(f)| \leq c\frac{\sigma(f)}{\sqrt{n}}\right) = \frac{1}{\sqrt{2\pi}} \int_{-c}^{c} e^{-x^2/2} \mathrm{d}x$$

for some constant c, which gives a probabilistic upper bound for the error which has a dimension-independent convergence rate of $O(n^{-1/2})$, and allows the computation of approximate confidence intervals (Niederreiter 1992, p.5).

MC is applicable to a wide range of problems as it only requires a continuous square-integrable function and a probablity measure from which we can sample. Even though MC has the advantage of exhibiting an error convergence rate of $O(n^{-1/2})$ that is independent from the number of dimensions d, it is often regarded as too slow (Lemieux 2009, pp.11-12). Considering the RMSE in equation (8), a higher accuracy can be achieved by either increasing n or, more effectively, decreasing the variance σ_f^2. For

this reason, numerous variance reduction techniques have been introduced such as importance sampling, stratified sampling or the use of control variates (see e.g. L'Ecuyer 1994). A second approach is to deploy alternative sampling mechanisms for which the convergence rate of the corresponding error is faster than for MC: quasi-Monte Carlo integration based on low-discrepancy points.

3.2 Quasi-Monte Carlo Methods

Similar to standard MC techniques, quasi-Monte Carlo (QMC) methods approximate integrals by an equal-weight cubature rule as in (7), but usually restrict the domain $\mathcal{X}$ to the d-dimensional half-open unit cube $[0,1)^d$. Instead of using IID pseudo-random numbers as evaluation points, deterministic low-discrepancy point sets are designed on the unit cube to cover this domain more evenly than random, guarantee error bounds and possibly achieve a faster convergence rate for sufficiently smooth functions (Morokoff et al. 1995). The fundamental difference between MC and QMC methods is best illustrated by Figure 1, which gives a comparison between a pseudo-random and a low-discrepancy point set (i.e. a Halton sequence) in $d = 2$. Note the pseudo-random points building clumps and voids since the different points do not 'know' about each other. This limiting factor for the accuracy of standard MC is remedied by using deterministic points with mutual correlations constructed to prevent clumping (Caflisch 1998).

A commonly used measure of discrepancy in the QMC context is the star discrepancy,

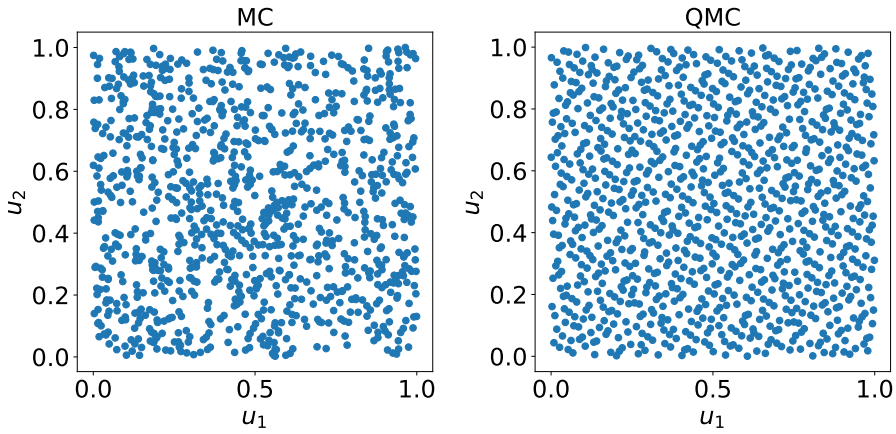


Figure 1: Comparison of 2^{10} MC and QMC points (Halton sequence)

which measures the irregularity of a finite point set $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n \in [0,1)^d$ (Niederreiter 1992, p.14):

$$D_n^* = D_n^*(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) = \sup_{\boldsymbol{a} \in [0,1)^d} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\boldsymbol{u}_i \in [0, \boldsymbol{a})} - \prod_{j=1}^{d} a_j \right|$$

where $\mathbb{1}_{\boldsymbol{u}_i \in [0,\boldsymbol{a})}$ is an indicator function with value 1 if $\boldsymbol{u}_i \in [0, \boldsymbol{a})$ and 0 if $\boldsymbol{u}_i \notin [0, \boldsymbol{a})$ and $0 \leq a_j < 1$. Thus, if D_n^* is small, the fraction of n points inside $[0, \boldsymbol{a})$ is close to the proportion of the unit cube taken up by the respective box. Using D_n^*, it is possible to define that an infinite sequence $\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots \in [0,1)^d$ is uniformly distributed modulo 1 (u.d. mod 1) if $D_n^*(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) \to 0$ as $n \to \infty$ (Niederreiter 1992, p.17). A u.d. mod 1 sequence of points in $[0,1)^d$ satisfies $\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\boldsymbol{u}_i \in [\boldsymbol{a}, \boldsymbol{b})} = \prod_{j=1}^{d} [a_j, b_j)$ for $n \to \infty$ and all d-dimensional intervals $[\mathbf{a}, \mathbf{b}) \subseteq [0,1)^d$ (Aistleitner et al. 2014). Considering the rate at which the star discrepancy decreases as $n \to \infty$, a finite sequence of points is regarded a low discrepancy sequence if

$$D_n^*(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) = O\left(n^{-1} (\log n)^d\right)$$

as $n \to \infty$. This rate can also be expressed as $O(n^{-1+\epsilon})$ for any $\epsilon > 0$ as $\log n$ is asymptotically negligible in comparison to any finite positive power of n (Dick et al. 2013). However, note that this finite power is dependent on dimension d.

As the random points in MC are in QMC replaced by deterministic ones, the LLN and CLT are no longer valid. For QMC, the following result ensures convergence of QMC methods (Kuipers et al. 1974): If the function f is a Riemann integrable and $\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots \in [0,1)^d$ is u.d. mod 1, then as $n \to \infty$

$$\left| \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{u}_i) - \int_{[0,1)^d} f(\mathbf{x}) \mathrm{d}\mathbf{x} \right| \to 0.$$

Thus, the QMC estimate $Q_{n,d}(f)$ will converge to $I_d(f)$ if a low-discrepancy set of points is used. To study the error, instead of using the CLT, QMC methods rely on the

Koksma-Hlawka inequality for $d \geq 1$ and $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n \in [0,1)^d$ (Caflisch 1998):

$$\left| \frac{1}{n} \sum_{i=1}^{m} f(\boldsymbol{u}_i) - \int_{[0,1)^d} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| \leq D_n^*(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) V_{HK}(f) \tag{9}$$

where V_{HK} denotes the total variation in the sense of Hardy and Krause. For a one-dimensional function with continuous first derivative the Hardy-Krause (HK) variation is defined as $V_{HK}(f) = \int_0^1 |\frac{\mathrm{d}f}{\mathrm{d}x}| \mathrm{d}u$ and for a d-dimensional function it is defined in terms of the integral of partial derivatives (Caflisch 1998):

$$V_{HK}(f) = \int_{[0,1]^d} \left| \frac{\partial^d f}{\partial u_1 \ldots \partial u_d} \right| \mathrm{d}u_1 \ldots \mathrm{d}u_d + \sum_{i=1}^{d} V_{HK}(f_1^{(i)})$$

where $f_1^{(i)}$ denotes the restriction of the function f to the boundary $u_i = 1$. These restrictions are formulated recursively as they are functions of $d-1$ variables (Caflisch 1998).

The inequality in (9) gives an upper bound on the error $|I_d(f) - Q_{n,d}(f)|$ using the product of a measure of non-uniformity of the sequence of evaluation points and a measure of the smoothness of f. This result holds with certainty as it is not probabilistic and holds for finite n (Caflisch 1998). However, both the star discrepancy and the HK variation are very hard to compute in practice (Gnewuch et al. 2009; Owen 2005) rendering the deterministic bound unusable. Nonetheless, the Koksma-Hlawka inequality gives an important result in QMC theory: if a low discrepancy sequence is used, $|I_d(f) - Q_{n,d}(f)| = O(n^{-1+\epsilon})$ for any $\epsilon > 0$ will be achieved. Thus, if $V_{HK}(f) < \infty$, QMC yields better accuracy then MC for large enough n, although there is no information on how large n has to be for the asymptotic rate to be relevant. The latter issue points at a drawback of the classical QMC theory: the error bound in equation (9) grows exponentially with dimension d implying that the methods are not useful for very high-dimensional problems.

Generally, QMC methods can be divided into two main families of construction approaches: *lattice rules* and *digital nets* and *sequences* (Dick et al. 2013). Examples from both families will be given in the following.

Examples of QMC Point Set Constructions

Lattice rules, as introduced by Korobov (1959), refer to an equal-weight cubature rule for which the cubature points in the half-open unit cube $[0,1)^d$ belong to an integration lattice. The integration lattice is a discrete subset of $\mathbb{R}^d$, which is not only closed under addition and subtraction, but also contains $\mathbb{Z}^d$ as a subset (Sloan et al. 1994). Generally, the origin $\mathbf{0}$ is included in every lattice point set and projecting these points onto each axis yields equally spaced points. For every lattice rule, there is a representation based on a multiple sum including generating vectors. The minimum number of required generating vectors determines the rank of the lattice rule, which can vary between one and d (Kuo et al. 2005).

The simplest example of this family of QMC methods is the n-point *rank-one lattice rule* in d dimensions (Kuo et al. 2006), which has the cubature points

$$\boldsymbol{u}_i = \left\{ \frac{iz}{n} \right\}, \quad i = 1, \ldots, n,$$

where $z \in \mathbb{Z}^d$ denotes the generating vector and the braces indicate that only the fractional part of a real number is used, i.e. $\{u\} := u - \lfloor u \rfloor$. The generating vector z is an n-dimensional integer vector, which has no factor in common with n. One possible approach to construct z is the Korobov construction (Sloan et al. 1994)

$$z = z(a) := (1, a, a^2, \ldots, a^{d-1}) \bmod n,$$

where a is an integer satisfying $1 \leq a \leq n-1$ and $\gcd(a, n) = 1$, i.e. the greatest common divisor of a and n is 1. The value of a minimising a desired error criterion can be found by searching through the $n-1$ possible choices. Further, lattice rules are constructed to work best for a prime number of samples (Sloan et al. 2003).

Digital nets and sequences are the second family of methods used to generate QMC points. Digital nets are based on the idea of subdividing the unit cube into intervals and position the points in a way that the number of points in an interval corresponds to the size and shape of this interval. To define a (t, m, d)-*net in base* b, with the integer parameters t, m, d and b, where d corresponds to the dimension of the space for $\boldsymbol{u}$,

we first introduce a specific subinterval of $[0,1)^d$, the elementary interval in base b for integers $d \geq 1$ and $b \geq 2$ (Owen 1997):

$$\prod_{j=1}^{d} \Big[\frac{a_j}{b^{k_j}}, \frac{a_j + 1}{b^{k_j}} \Big).$$

a_j and k_j are integers with $0 \leq a_j \leq b^{k_j}$ and $k_j \geq 0$, where $k_1 + k_2 + \cdots + k_d = m - t$, $m \geq 1$ and $t \leq m$. Then, a (t, m, d)-net in base b is a set of b^m points in $[0,1)^d$ such that every elementary interval contains exactly b^t of these points. While the (t, m, d)-net is a finite sequence, an extensible version can be constructed by concatenating an infinite sequence of (t, m, d)-nets for any $m \geq t$ (Niederreiter 1987). Well-known examples of this approach include the Sobol' sequence (Sobol' 1967) or the Faure sequence (Faure 1951), which both work well with sample sizes that are powers of 2. Details on the construction of digital nets and sequences can be found for instance in Dick et al. (2010).

3.3 Randomised Quasi-Monte Carlo Methods

Although QMC techniques can possibly yield faster convergence rates in comparison to standard MC methods, the resulting estimators are biased and error estimates cannot be provided (Caflisch 1998). Randomised quasi-Monte Carlo (RQMC) is designed to overcome this drawback by introducing randomisation to the QMC approach while preserving the underlying properties of the QMC point distribution, thereby providing an unbiased estimator, error estimates and faster convergence rates than MC methods for smooth functions (Dick et al. 2013). The difference between a QMC (i.e. a Sobol' sequence) and an RQMC point set (i.e. a scrambled Sobol' sequence) is illustrated for $d = 2$ in Figure 2. It is apparent that after randomising the QMC points the domain is still covered evenly without any clumps or voids.

Generally, RQMC methods require the points $\boldsymbol{u}_1^{RQMC}, \ldots, \boldsymbol{u}_n^{RQMC}$ to be individually uniformly distributed on $[0,1)^d$, but, as a set, of low-discrepancy. The RQMC point construction applies randomisations to the QMC points $\boldsymbol{u}_1^{QMC}, \ldots, \boldsymbol{u}_n^{QMC} \in [0,1)^d$ to generate the RQMC points $\boldsymbol{u}_i^{RQMC} \sim U([0,1)^d)$, while partly preserving the QMC structure from $\boldsymbol{u}_i^{QMC}$ in the new cubature points $\boldsymbol{u}_i^{RQMC}$ (Hickernell 1996). Due to the randomisation, the usual equal-weight cubature rule from (7) provides an unbiased

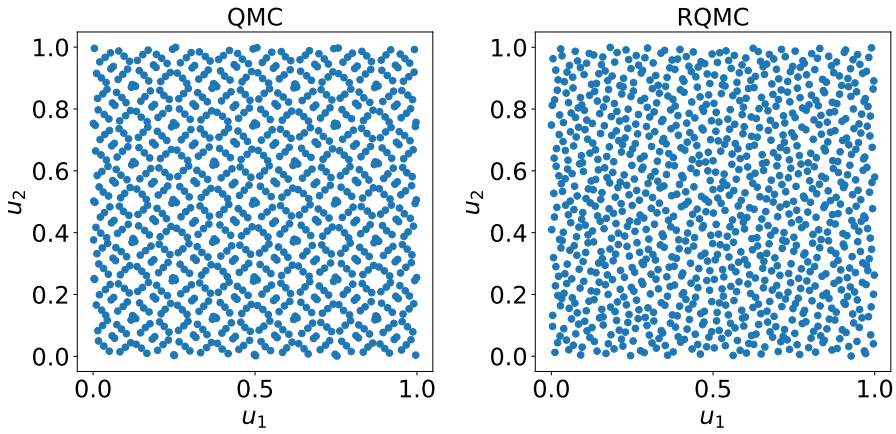


Figure 2: Comparison of 2^{10} QMC (Sobol' sequence) and RQMC points (scrambled Sobol' sequence)

estimator for the considered integral. Further, the QMC error bounds in (9) apply, so that the RMSE is of order $O(n^{-1+\epsilon})$ for any $\epsilon > 0$, implying an asymptotically better accuracy than MC methods if $V_{HK}(f) < \infty$ (L'Ecuyer et al. 2008).

Independently applying randomisations to $\boldsymbol{u}_1^{QMC}, \ldots, \boldsymbol{u}_n^{QMC}$ R times allows R approximations of the integral. Averaging these $Q_{n,d}^{(1)}(f), \ldots, Q_{n,d}^{(R)}(f)$, gives the final integral approximation $\bar{Q}_{n,d,R}(f)$ (Dick et al. 2013):

$$\bar{Q}_{n,d,R}(f) = \frac{1}{R} \sum_{r=1}^{R} Q_{n,d}^{(r)}(f) \tag{10}$$

for which $\mathbb{E}[\bar{Q}_{n,d,R}(f)] = I_d(f)$ since $Q_{n,d}^{(1)}(f), \ldots, Q_{n,d}^{(R)}(f)$ are IID. An unbiased variance estimate is given by

$$\widehat{\mathrm{Var}}(\bar{Q}_{n,d,R}(f)) = \frac{1}{R(R-1)} \sum_{r=1}^{R} \left(Q_{n,d}^{(r)} - \bar{Q}_{n,d,R}(f) \right)^2 \tag{11}$$

enabling the computation of confidence intervals for $\bar{Q}_{n,d,R}(f)$.

Examples of Randomisation Techniques

Randomisation techniques for RQMC, which preserve the properties of the QMC point sets, are available for both lattice rules and digital nets and sequences.

Shifting is utilised for lattice rules to obtain so-called shifted lattice rules (Kuo et al.

2005). The basic idea is to shift all points by the same amount into the same direction and to wrap all points, which fall outside the unit cube on one side, back inside from the opposite side. The shift is a vector $\boldsymbol{\Delta}_r$, which is randomly sampled from the uniform distribution on $[0,1)^d$, and yields the RQMC point set $\boldsymbol{u}_1^{RQMC}, \ldots, \boldsymbol{u}_n^{RQMC}$ by shifting the QMC points $\boldsymbol{u}_1^{QMC}, \ldots, \boldsymbol{u}_n^{QMC}$ by $\boldsymbol{\Delta}_r$, i.e. $\boldsymbol{u}_i^{RQMC} = \{\boldsymbol{u}_i^{QMC} + \boldsymbol{\Delta_r}\}$ for $i = 1, \ldots, n$ (Kuo et al. 2005).

Continuing the example of the rank-one lattice rule introduced in section 3.2, the more general *shifted rank-one lattice rule* has the cubature points

$$\boldsymbol{u}_i^{RQMC} = \left\{ \frac{i\boldsymbol{z}}{n} + \Delta \right\}, \quad i = 1, \ldots, n \tag{12}$$

where $\boldsymbol{z}$ is the generating vector and $\boldsymbol{\Delta} \in [0,1)^d$ is the shift. The preferred sample sizes are still primes (Kuo 2003).

Randomisation techniques for digital nets and sequences are required to preserve the corresponding relevant structure implying that a (t, m, d)-net should, with probability 1, still be a (t, m, d)-net after introducing randomness. *Scrambling* satisfies this requirement: it is based on the idea of splitting a unit cube filled with points of a (t, m, d)-net in base b into b equally-sized slabs $[\ell/b, (\ell+1)/b) \times [0,1)^{d-1}$ for $\ell = 0, 1, \ldots, b-1$ and then randomly permuting them. The resulting point positions would still correspond to a (t, m, d)-net (Dick et al. 2013).

A well-known example is the *nested uniform scramble* introduced by Owen (1997). The algorithm continues to scramble the points within each slab b and then within each sub-slab $b^2, b^3, \ldots$ until the slabs are too thin to have an effect on the floating point representation of the point u_{ij}^{RQMC}. All coordinate dimensions are shuffled independently according to this pattern. Mathematically, this algorithm is a permutation operation on the base b digits of a point u_{ij}^{QMC}:

$$u_{ij}^{QMC} = \sum_{k=1}^{K} u_{ijk}^{QMC} b^{-k} \in [0,1) \quad \rightarrow \quad u_{ij}^{RQMC} = \sum_{k=1}^{K} u_{ijk}^{RQMC} b^{-k} \tag{13}$$

where u_{ijk}^{QMC} and u_{ijk}^{RQMC} denote the k-th digits of the points $u_{i,j}^{QMC}$ and $u_{i,j}^{RQMC}$ respectively. The required digits $u_{ijk}^{RQMC} \in \{1, \ldots, b\}$ are determined by the mapping

$u_{ijk}^{QMC} \to u_{ijk}^{RQMC} = \pi_{jk}(u_{ijk}^{QMC})$, where π_{jk} is a random permutation of $\{1, \ldots, b\}$ dependent on the previous digits $u_{ij1}^{QMC}, \ldots, u_{ij,k-1}^{QMC}$. Permutations with different indices are assumed to be chosen mutually independent from each other and each permutation has equal probability. Using the nested uniform scramble, the RMSE of a (λ, t, m, d)-net for a smooth function f and sequences of values $n = \lambda b^m$ is $O(n^{-3/2+\epsilon})$ which is smaller than the QMC rate of convergence (Owen 1997).

3.4 Practical Issues

The effect of using these low-discrepancy point sets for integration problems is illustrated by following simple example. Consider the integral $f(x) = \int_{-1}^{1} x^2 \mathbb{U}(\mathrm{d}x)$ where the probability measure $\mathbb{U}$ corresponds to the uniform distribution on $[0, 1)$. The analytical solution to this problem is known to be $\frac{2}{3}$. Using a set of points on $[0, 1)$, $u_1, \ldots, u_n$, the integral can be approximated by $f(x) \approx \frac{b-a}{n} \sum_{i=1}^{n} g(a+(b-a)u_i) = \frac{2}{n} \sum_{i=1}^{n} (-1+2u_i)^2$. Figure 3 depicts the simulation results for the integral approximation via MC, QMC and RQMC approaches dependent on the number of samples used. For MC and RQMC, $R = 10$ repetitions were averaged for the final approximation.

It is obvious that all methods are able to converge to the true value of the integral, but the MC approximation continues to fluctuate considerably around this value as the

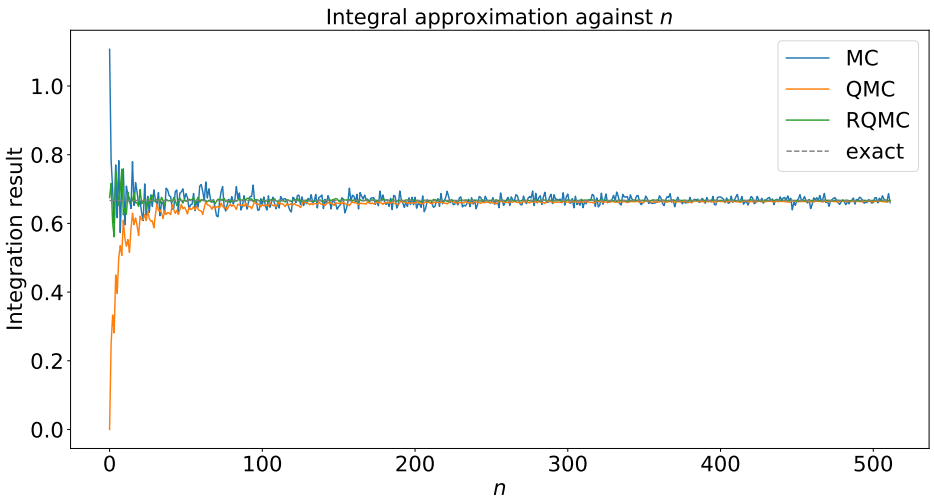


Figure 3: Approximation of the integral $f(x) = \int_{-1}^{1} x^2 \mathbb{U}(\mathrm{d}x)$ using MC, QMC (Sobol' sequence) and RQMC (scrambled Sobol' sequence) methods

number of samples increases. Plotting the absolute integration error for this problem against the number of samples in Figure 4 reveals the faster convergence rates of QMC and RQMC in comparison to MC for larger n. For smaller n, the constant $(\log n)^d$ in the convergence rate of the QMC error with $d = 1$ in this case seems to dominate the results for the QMC and RQMC approximations. In this simple example, it only takes about 150 samples for QMC to achieve MC precision. In high-dimensional settings, however, we have to keep in mind that a much larger number of samples might be necessary to achieve faster convergence than MC as the rate of $O(n^{-1+\epsilon})$ for $\epsilon < 0$ will only hold in the limit.

Another issue that remains in practice is the transformation to other distributions, since QMC and RQMC point sets are usually defined on the unit cube. It is desirable to convert the low-discrepancy point sets in a way, in which their structure and, thus, their desirable properties which make quasi-Monte Carlo converge faster than regular Monte Carlo, are conserved. In many applications, for example, points need to be constructed to mimic a (multivariate) Gaussian distribution. There are various possibilities to achieve this such as using the inverse transform (Pillards et al. 2006), the Box-Muller transform (Box et al. 1958) or Marsaglia's polar method (Marsaglia et al. 1964). In this thesis, we will specifically consider the Box-Muller transform to obtain standard

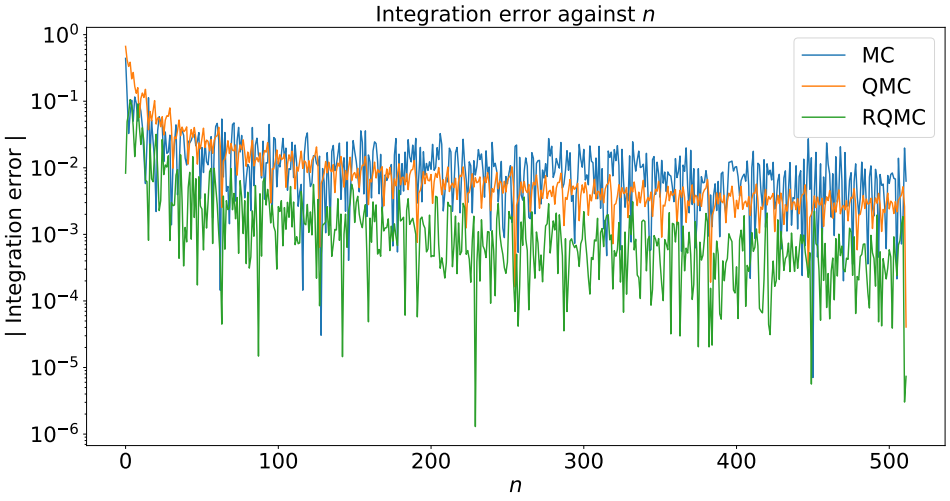


Figure 4: Integration error of the integral $f(x) = \int_{-1}^{1} x^2 \mathbb{U}(\mathrm{d}x)$ using MC, QMC (Sobol' sequence) and RQMC (scrambled Sobol' sequence) methods

normally distributed samples z_i from points $u_i \in [0, 1)$ for $i = 1, \ldots, n$ (Box et al. 1958):

$$z_1 = \sqrt{-2\ln(u_1)}\cos(2\pi u_2)$$
$$z_2 = \sqrt{-2\ln(u_1)}\sin(2\pi u_2).$$

Ökten et al. (2011) prove that this Box-Muller transform preserves the low-discrepancy structure and is, therefore, a valid choice to generate low-discrepancy sequences from the normal distribution.

In the context of transforming uniform point sets to normally distributed ones, a further problem arises: QMC point sets usually include the origin $\mathbf{0}$. When using the Box-Muller transform, trying to calculate $\log 0$ will make the first Gaussian point infinite. Owen (2020) discusses two possibilities to address this problem. The first option is to simply drop the first point, whereas the second option is to apply a single randomisation to the QMC point set which will avoid the boundaries of the unit cube. Owen (2020) finds the first option to be detrimental for the convergence rate, so that we adopt the second option for our numerical experiments. Thus, whenever a QMC point set is mentioned, we in fact refer to a fixed but once randomised QMC point set.

4 A Novel Application for QMC Methods: Inference for Generative Models

In modern statistics, the inference tools are required to scale well with model complexity and the size of datasets. This especially holds true for intractable generative models which are inherently complex as their likelihood or an approximation thereof is not necessarily available (Mohamed et al. 2016). As a standard approach, inference methods, and minimum divergence estimators for generative models in particular, typically rely on MC methods to obtain estimates (Bassetti et al. 2006; Genevay et al. 2018; Briol et al. 2019). As discussed extensively in section 3.1, MC methods are suitable for a wide range of applications and even high-dimensional settings due to an error convergence rate independent of dimension. However, the slow speed of convergence of $O(n^{-1/2})$ for large n is the reason for MC-based inference tools failing to scale well with data size.

This has motivated the attempt to leverage the faster error convergence rates of QMC and RQMC in various applications: it is extensively used in computational finance for the valuation of complex derivatives where the method was found to yield faster and smoother convergence as well as higher accuracy even in high-dimensional settings, see for instance L'Ecuyer (2009) for an overview. Boyle et al. (2005) successfully use QMC to not only estimate the prices of options, but also partial first and second derivatives of these prices. QMC and RQMC methods were further applied in the context of parameter estimation via ML using Newton–Raphson (Pan et al. 2007) or expectation–maximization algorithms (Jank 2005).

These successful applications provide the motivation to explore the use of QMC methods for the class of minimum divergence estimators for generative models. We concentrate on one specific instance, the minimum MMD framework proposed by Briol et al. (2019), and attempt to improve the scalability of these estimators. In section 2.3.1, we touched upon the possible starting point for modifying the estimators: using MC integration, Briol et al. (2019) derive the rate at which an empirical measure $\mathbb{P}^n$ converges to the true probability measure $\mathbb{P}$ in terms of their MMD in the concentration bound in equation (5). This bound decreases as $n^{-1/2}$ which equals the MC convergence rate. Thus, it seems promising to utilise QMC and RQMC integration to leverage their faster convergence rates of $O(n^{-1+\epsilon})$ for any $\epsilon > 0$ to gain the higher efficiency needed for the estimator to scale better with data size.

To do so, the optimisation algorithms presented in section 2.3.1 has to be modified. After initialising at $\hat{\theta}^{(0)} \in \Theta$, the first of the two steps over which both SGD and NSGD algorithms iterate has to be adjusted depending on whether the MC, QMC or RQMC method is used. To isolate the effect of the point set choice, the simulation of samples from the generative model of interest using the MC approach is changed to

1a. sample $\{\mathbf{u}_i^{MC}\}_{i=1}^n \overset{\text{IID}}{\sim} U([0,1)^d)$, transform the $\mathbf{u}_i^{MC}$ into $\mathbb{U}$-distributed samples if necessary, and compute $\mathbf{x}_i = G_{\hat{\theta}^{(t-1)}}(\mathbf{u}_i^{MC})$ for $i = 1, \ldots, n$.

Thus, instead of sampling from the measure $\mathbb{U}$ directly we sample uniformly distributed points and transform them into the required probability distribution if $\mathbb{U}$ is different from the uniform distribution over $[0,1)^d$. In case of QMC, we construct a single low-discrepancy point set and re-use it at every descent step:

1b. get the pre-defined QMC points $\{\mathbf{u}_i^{QMC}\}_{i=1}^n \in [0,1)^d$, transform the $\mathbf{u}_i^{QMC}$ into $\mathbb{U}$-distributed samples if necessary, and compute $\mathbf{x}_i = G_{\hat{\theta}^{(t-1)}}(\mathbf{u}_i^{QMC})$ for $i = 1, \ldots, n$.

Again, the QMC point set is transformed into the desired probability distribution in case the measure $\mathbb{U}$ is different from the uniform distribution over $[0,1]^d$. For the RQMC approach, the only difference to the aforementioned procedure is that the pre-defined QMC point set is randomised at every iteration:

1c. randomise the QMC points $\{\mathbf{u}_i^{QMC}\}_{i=1}^n \in [0,1)^d$, transform the resulting $\mathbf{u}_i^{RQMC}$ into $\mathbb{U}$-distributed samples if necessary, and compute $\mathbf{x}_i = G_{\hat{\theta}^{(t-1)}}(\mathbf{u}_i^{RQMC})$ for $i = 1, \ldots, n$.

The second step of the iterative optimisation algorithm is similar to step 2 in section 2.3.1 and depends on whether SGD or NSGD is deployed. This first method-depending step will also be utilised for the simulation of samples to investigate the effect of QMC and RQMC methods on two main competitors of the MMD in the context of inference for generative models, namely the 1-Wasserstein distance and the Sinkhorn loss of Genevay et al. (2018).

We would expect the modified algorithms based on QMC and RQMC to benefit from the faster convergence rates of these methods as compared to MC, thereby improving the scalability of the minimum MMD estimators considerably. These faster convergence rates should translate into a similar accuracy of QMC while using less data points. The assumed increase in efficiency should therefore come at negligible additional computational cost arising from the construction of QMC and RQMC point sets as the most costly operations in numerical integration are the function evaluations (Giles et al. 2008). Further, these improvements are supposed to be easy to implement in practice as libraries for the construction of QMC and RQMC point sets are readily available for most commonly used programming languages such as R (Hofert et al. 2016), MATLAB (Choi et al. 2020a) or Python (Choi et al. 2020b).

However, as discussed in section 3.2, the improved error convergence rate of QMC and RQMC methods is also associated with drawbacks: compared to MC stronger conditions on the integrand are required, i.e. the integrand has to be of bounded HK variation (Caflisch 1998). Moreover, it is only possible to obtain point sets on the unit cube,

so that an additional transformation into the desired distribution is usually necessary in order to obtain samples from the generative model of interest. The possible sample sizes are also determined by the chosen QMC point construction which typically only achieve a low discrepancy for a certain number of points such as powers of two for Sobol' sequences (Sobol' 1967) or prime numbers for lattice rules (Sloan et al. 2003). Using standard QMC methods, we might observe a dimension-dependent performance since the error convergence rates only hold asymptotically implying that we do not know which sample size is necessary for the algorithm to be more efficient than when using MC (Drew et al. 2006).

The following section empirically investigates how the modified minimum MMD estimators based on QMC and RQMC methods compare to the MC-based version.

5 Numerical Experiments

The numerical experiments cover different generative models with various data and parameter dimensions to investigate the effect of using low-discrepancy points on the minimum MMD estimator. In the M-closed setting, insights into the impact of QMC and RQMC methods in terms of precision and convergence under varying circumstances are provided. Unless stated differently, Halton sequences and scrambled Halton sequences were deployed as QMC and RQMC point sets respectively. For their implementation, the Python library `QMCPy` (Choi et al. 2020b) was used. The kernel is chosen to be the Gaussian kernel as presented in section 2.1 with the lengthscale being determined using the *median heuristic*. The median heuristic was proposed by Gretton et al. (2012) for RBF kernels and sets the lengthscale proportionally to the median of all pairwise distances in the aggregated sample, i.e. $l = \sqrt{\text{median}(\|\mathbf{y}_i - \mathbf{y}_j\|_2^2/2)}$ for $1 \leq i, j \leq m$ (Muandet et al. 2014). Reproducible code can be found in this GitHub repository[1].

5.1 Gaussian Location Model

The first model we will investigate is the *Gaussian location model*. It describes a d-dimensional isotropic Gaussian distribution $N(\theta, \sigma^2 I_{d\times d})$ for which the standard devi-

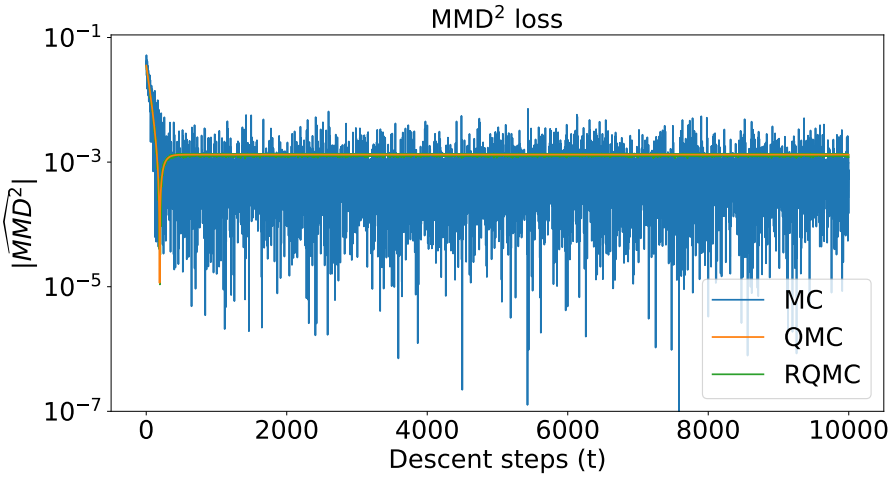
[1] www.github.com/johannnamr/MSc-Project-Code

ation $\sigma > 0$ is known but the mean $\theta \in \mathbb{R}^d$ is not. This implies that the parameter dimension equals the data dimension, i.e. $p = d$. For $\mathbf{z}_i \in \mathbb{R}^d$ and $\mathbf{u}_i \in [0,1)^d$, the generator can be defined as $G_\theta(\mathbf{u}_i) = \mathbf{z}_i(\mathbf{u}_i) + \theta$ where $z_{i,2k-1} = \sqrt{-2\log u_{i,2k-1}}\cos(2\pi u_{i,2k})$ and $z_{i,2k} = \sqrt{-2\log u_{i,2k-1}}\sin(2\pi u_{i,2k})$ for $k = 1, \ldots, d/2$ if d is even and $k = 1, \ldots, (d+1)/2$ with $z_{i,2(d+1)/2}$ being discarded if d is odd. Thus, we have $G_\theta : [0,1)^d \to \mathbb{R}^d$ if d is even and $G_\theta : [0,1)^{d+1} \to \mathbb{R}^d$ if d is odd. In case of MC, the $\mathbf{u}_i$ are IID draws from $U([0,1)^d)$ if d is even or $U([0,1)^{d+1})$ if d is odd, whereas, for QMC and RQMC, the $\mathbf{u}_i$ are deterministic low-discrepancy points set on $[0,1)^d$ or $[0,1)^{d+1}$. The partial derivative of this generator is given by $\nabla_\theta G_\theta(\mathbf{u}_i) = I_{d\times d}$. We consider a data set consisting of samples $\{\mathbf{y}_j\}_{j=1}^m \overset{\text{IID}}{\sim} \mathbb{Q}$ where $\mathbb{Q} = \mathbb{P}_{\theta^*}$ and θ^* is the true parameter. For the following experiments, σ is set to 2. Clearly, many quantities of interest could be calculated in closed form and thus the model does not call for advanced optimisation algorithms, however, this very simple model is a good starting point to analyse the effects of using low-discrepancy points for the simulation of data in the minimum MMD framework for varying dimensions. In all following experiments, the true parameter is set to $\theta^* = (\theta_1^*, \ldots, \theta_p^*)$ with $\theta_i^* = 1$ for $i = 1, \ldots, p$.

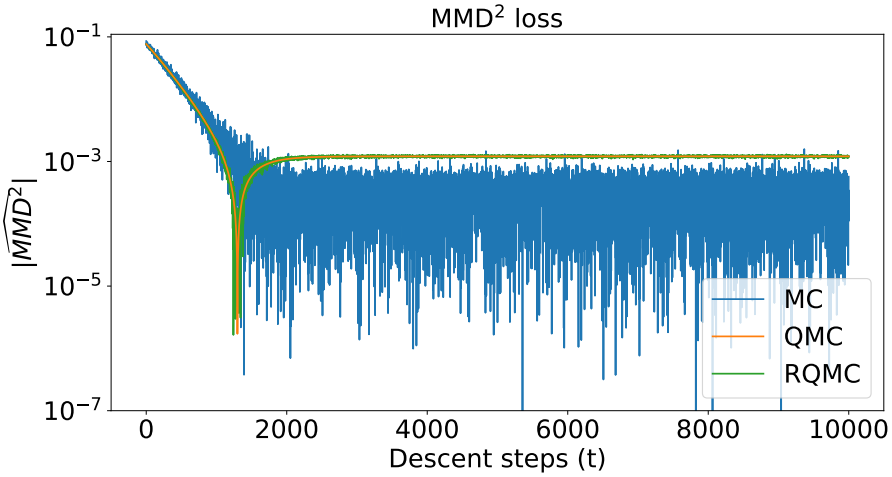
Optimisation

We begin by presenting the optimisation results when using SGD in a setting where the true parameter to recover is used to generate a large set of $m = 2^{11}$ samples from $\mathbb{Q}$. The optimisation algorithm is run for 10,000 iterations, while in each iteration $n = 2^9$ samples are generated using MC, QMC or RQMC point sets. The step size is fixed to 0.1, which is chosen to ensure convergence, and the start value for the SGD optimisation is set to $\theta_i^{(0)} = 0$ for $i = 1, \ldots, p$.

Figure 5 depicts $\widehat{\text{MMD}}^2(\mathbb{P}^n_{\hat{\theta}^{(t)}} \| \mathbb{P}^m_{\theta^*})$ against the number of iterations for $d = 1$ (5a) and $d = 10$ (5b) when using SGD for the optimisation. Generally, the algorithm converges for all methods and the loss for the QMC and RQMC approaches develops so similarly that the lines are widely overlapping. For MC, stochastic fluctuations are clearly visible. The squared MMD using the RQMC approach appears to fluctuate slightly more than for QMC, which is expected due to the randomisation of the low-discrepancy points at every iteration. Comparing the results for $d = 1$ and $d = 10$, it is apparent that for



(a) $d = 1$



(b) $d = 10$

Figure 5: MMD loss when optimising using SGD

higher dimensions, i.e. with increasing model complexity, more iterations of the SGD algorithm are needed to achieve convergence. This is further illustrated by Figure 18 in appendix B.1, which shows for $d = 5$ a convergence speed of the MMD-based loss between the ones observed for $d = 1$ and $d = 10$. In all three considered dimensions, QMC and RQMC appear to converge faster over the first descent steps, but then exhibit a converged absolute loss that is higher than MC due to the approximation of the squared MMD turning negative.

To measure the precision of the MMD estimator using SGD, we consider the mean squared error defined as $\text{MSE}^{(t)} = \frac{1}{t}\sum_{i=1}^{t}(\mathbf{y}_i - \mathbf{x}_i)^2$ with t being the number of performed

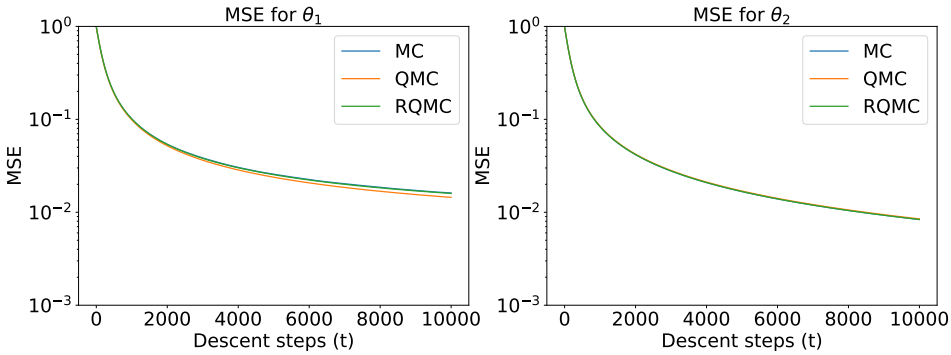


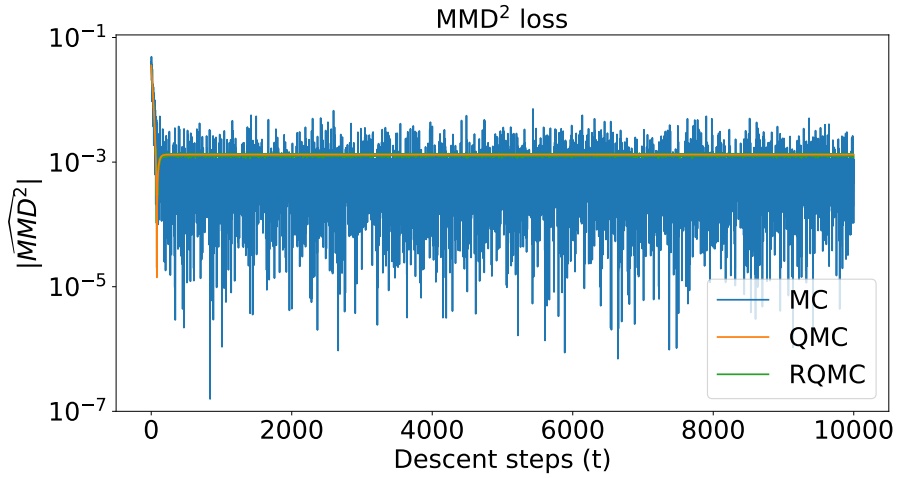
Figure 6: MSE when optimising using SGD for $d = 2$

descent steps. Figure 6 shows the MSE against the number of descent steps when $d = 2$ for both estimates $\hat{\theta}_1$ and $\hat{\theta}_2$. MC, QMC and RQMC methods are found to have similar precision for the Gaussian location model as all lines are overlapping. The same conclusion can be drawn for other considered dimensions (results not reported here).

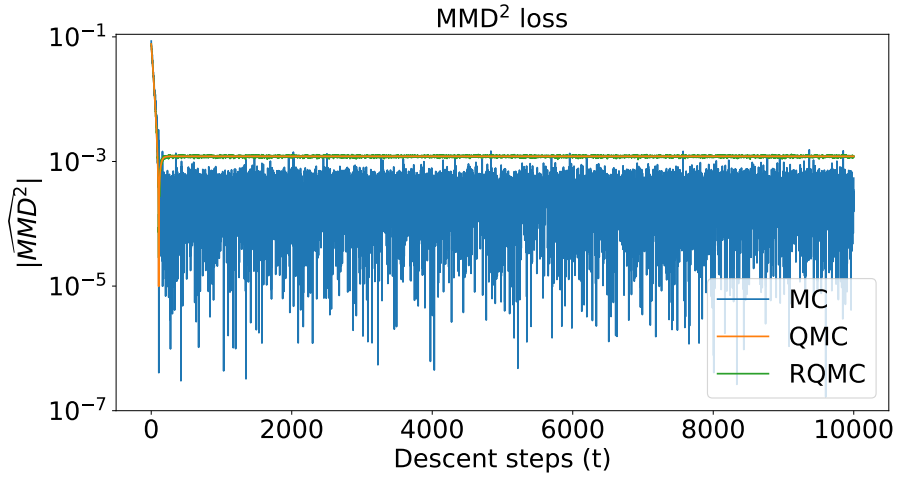
To compare the results for SGD with NSGD, we use the exact same setup as before but adjust the step size to 0.01 to ensure convergence. This is to accommodate that NSGD takes the geometry of the MMD-based loss function into account and thus moves more quickly into the direction of the optimum.

Again, we look at $\widehat{\text{MMD}}^2(\mathbb{P}^n_{\hat{\theta}(t)} \| \mathbb{P}^m_{\theta^*})$ plotted against the number of iterations for $d = 1$ and $d = 10$ in Figure 7. In general, we can state observations similar to those mentioned for SGD: MC, QMC and RQMC approaches all converge and most stochastic variation can be observed for MC. However, it is striking that the speed of convergence when using NSGD is much quicker for $d = 10$ than when using SGD (Figure 5a). The same behaviour can be reported for $d = 5$ in Figure 19 in appendix B.1. In all three considered dimensions, the absolute loss first converges faster for QMC and RQMC, but then exhibits a larger absolute converged loss for QMC and RQMC as the approximations turn negative. Thus, we can conclude that the MMD geometry can also be leveraged when using QMC and RQMC methods.

For $d = 2$, the precision in terms of the MSE against the number of iterations when using NSGD is illustrated in Figure 8. Again, a similar precision for MC, QMC and RQMC approaches can be noted. As expected from the previous results, the MSE decreases at a faster rate when optimising using NSGD instead of SGD (cf. Figure 6). In



(a) $d = 1$



(b) $d = 10$

Figure 7: MMD loss when optimising using NSGD

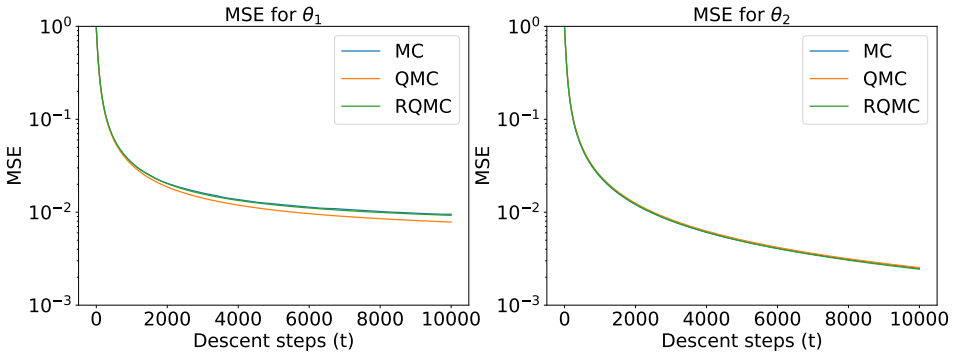


Figure 8: MSE when optimising using NSGD for $d = 2$

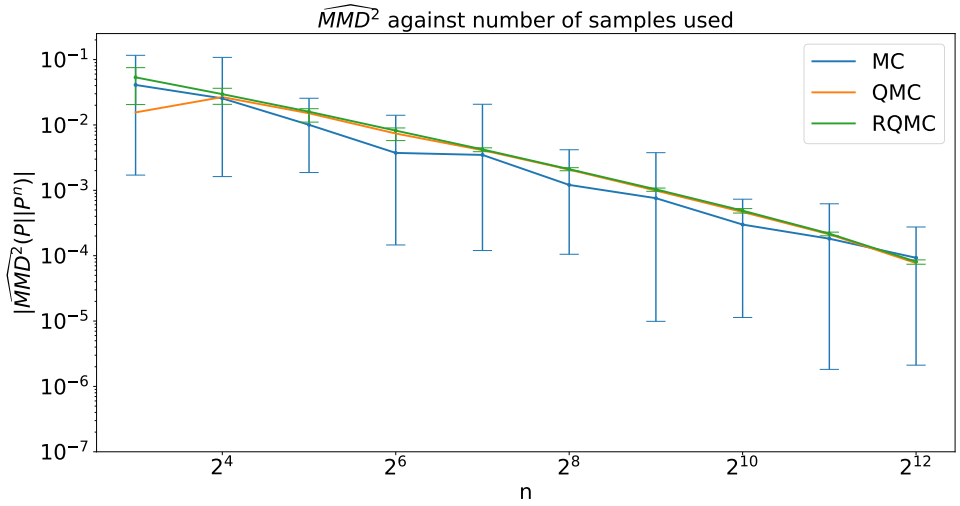
summary, for the Gaussian location model, there are no advantages in terms of precision visible for either sampling method regardless of whether SGD or NSGD is used.

Convergence

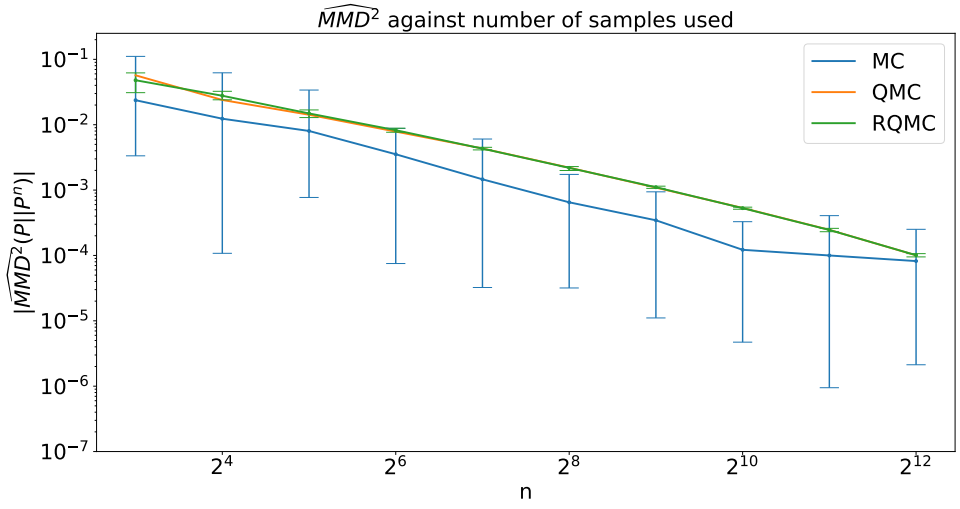
In order to analyse whether the data simulation in the MMD framework can be made more efficient using QMC and RQMC, it is examined how fast the empirical measure $\mathbb{P}_{\theta^*}^n$ converges to the true measure $\mathbb{P}_{\theta^*}$ in terms of their squared MMD. The true measure $\mathbb{P}_{\theta^*}$ is approximated using $\mathbb{P}_{\theta^*}^m$ with m large, i.e. $m = 2^{13}$. For this purpose, we plot the absolute $\widehat{\text{MMD}}^2(\mathbb{P}^n \| \mathbb{P}^m)$ against a range of different n for all methods. Following the approach of Cambou et al. (2017), the analysis is based on $R = 25$ repetitions per evaluated number of samples for MC and RQMC. The maximum and minimum values among these repetitions are given as error bars for the respective method.

Figure 9 presents the results of this experiment for $d = 1$ (9a) and $d = 4$ (9b). In both settings, the convergence speed of QMC and RQMC is very similar, but slower than the one of the MC approach for the whole considered range of n in case of $d = 4$ and for $n \leq 2^{11}$ in case of $d = 1$. Comparing the settings $d = 1$ and $d = 4$, this difference seems to be reinforced by an increasing dimension. Figure 9a further demonstrates that for large enough $n > 2^{11}$ the convergence rates achieved using QMC and RQMC points catch up with the one obtained by MC implying that, for a large enough number of samples, QMC and RQMC are more efficient in the case of $d = 1$. Figure 20 in appendix B.1 confirms that for $d = 3$ and a larger sample size of $n \geq 2^{12}$ this observation still holds true. The similarity of the patterns in Figure 9 for $d = 1$ and $d = 4$ suggests that this result might also be found in higher dimensions when even larger n are considered.

Generally, all these results could be an indicator for the constant of the convergence rate of the QMC and RQMC error $(\log n)^d$ influencing the speed at which the empirical measure $\mathbb{P}_{\theta^*}^n$ converges to the true measure $\mathbb{P}_{\theta^*}$ in terms of their squared MMD. Moreover, for both $d = 1$ and $d = 4$, the error bars illustrate that while the results for RQMC have the favourable property of becoming more stable with increasing n, the MC approach is subject to considerable variation.



(a) $d = 1$



(b) $d = 4$

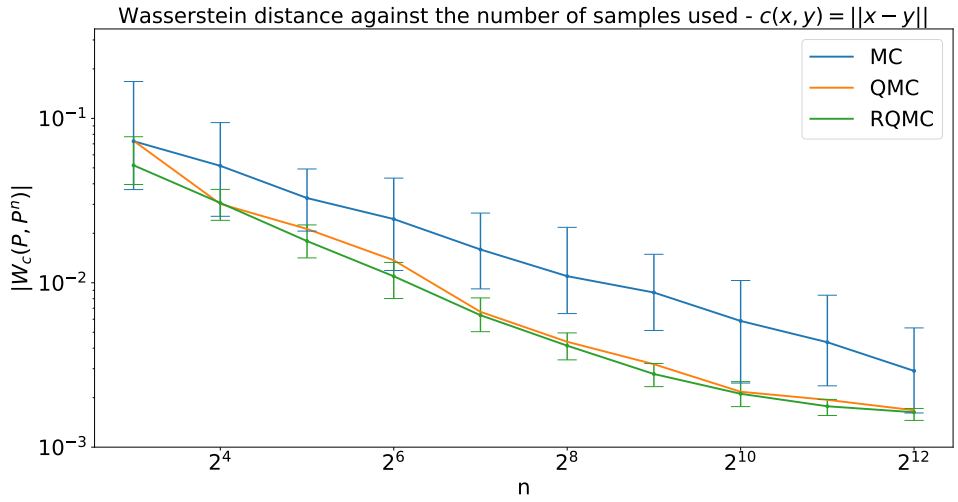
Figure 9: Convergence of $\mathbb{P}^n_{\theta^*}$ to $\mathbb{P}^m_{\theta^*}$ in terms of their squared MMD with $m = 2^{13}$

Comparison to Wasserstein Distance and Sinkhorn Loss

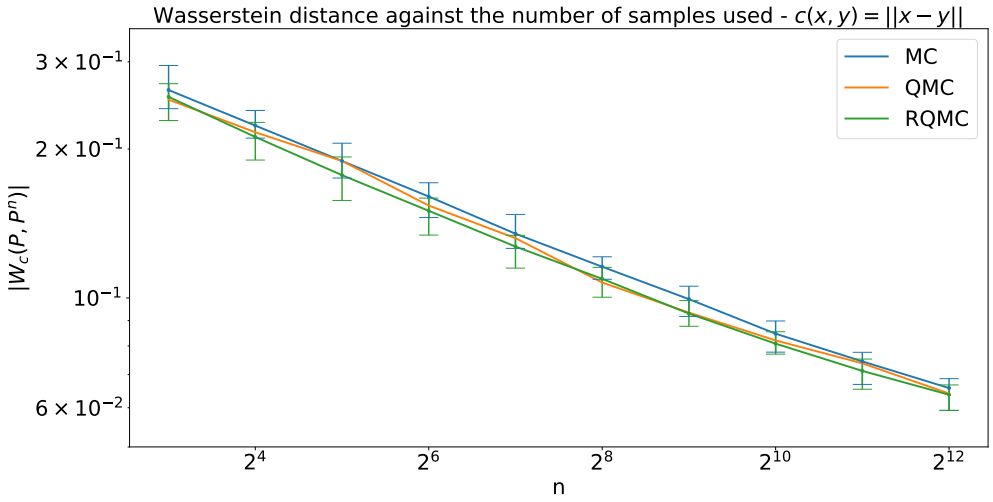
Having introduced a main competitor for the minimum MMD estimator in section 2.3.2, an estimator based on the Sinkhorn loss, the results described above can be compared with the convergence speed of an empirical measure $\mathbb{P}^n_{\theta^*}$ to the approximated true measure $\mathbb{P}^m_{\theta^*}$ in terms of their 1-Wasserstein distance as well as their Sinkhorn loss with squared ℓ^2 cost. The choice of cost in the form of $\|\cdot\|_2^2$ for the Sinkhorn loss is the baseline used for the numerical experiments by Genevay et al. (2018). For the approx-

imation of the true measure, we again set $m = 2^{13}$. Further, we again consider the average of $R = 25$ repetitions for MC and QMC at every evaluated number of samples n as in Cambou et al. (2017) and report the minimum and maximum of these repetitions in form of error bars.

Figure 10 shows the result of this experiment for the 1-Wasserstein distance, i.e. $W_1(\mathbb{P}^n_{\theta^*}, \mathbb{P}^m_{\theta^*})$ against n. For $d = 1$ (10a), the QMC and RQMC approaches converge considerably faster than when using MC even when considering the maximum values



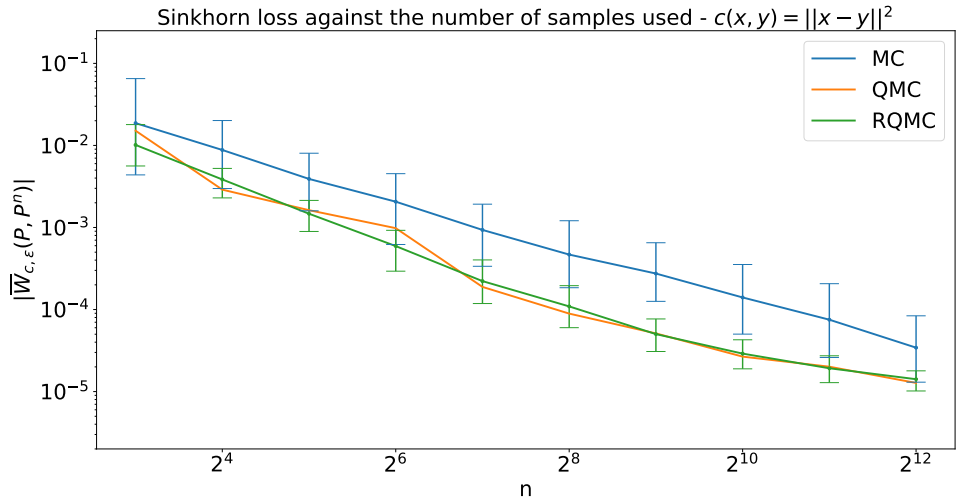
(a) $d = 1$



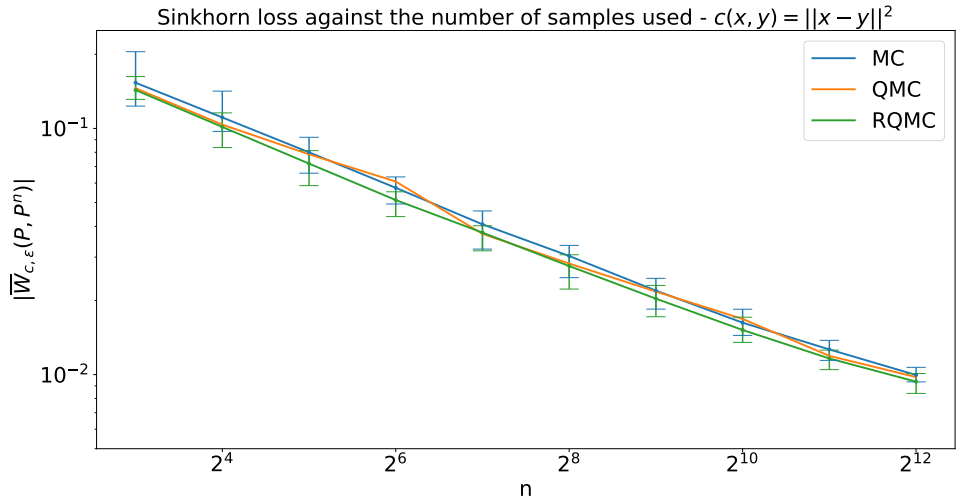
(b) $d = 5$

Figure 10: Convergence of $\mathbb{P}^n_{\theta^*}$ to $\mathbb{P}^m_{\theta^*}$ in terms of their 1-Wasserstein distance

for RQMC and the minimum values for MC over all repetitions for a given sample size. Increasing the dimension to $d = 5$ (10b) gives a similar convergence speed for all methods over all considered sample sizes. In line with the theoretical results presented in section 2.2.2, the 1-Wasserstein distance clearly suffers from the curse of dimensionality as the convergence rate of MC slows down considerably when switching from $d = 1$ to $d = 5$. This is a major difference to the behaviour of the MMD-based loss presented



(a) $d = 1$



(b) $d = 5$

Figure 11: Convergence of $\mathbb{P}^n_{\theta^*}$ to $\mathbb{P}^m_{\theta^*}$ in terms of their Sinkhorn loss with squared ℓ^2 cost and $\epsilon = 10^{-3}$

earlier. Additionally, for QMC and RQMC in $d = 5$, we can assume that the constant term in their error convergence rates dominates since there is no longer an advantage of these approaches over MC visible for the considered range of n. However, it cannot be determined from this analysis whether there is a larger n for which one method converges faster than the other.

In the same setting as for the 1-Wasserstein distance, Figure 11 depicts the results for the Sinkhorn loss based on the squared ℓ^2 cost and $\epsilon = 10^{-3}$, i.e. $\overline{W}_{\ell^2,\epsilon}(\mathbb{P}^n_{\theta^*}, \mathbb{P}^m_{\theta^*})$ against n. A similar pattern to the 1-Wasserstein distance can be noted: for $d = 1$ (11a), QMC and RQMC show a faster convergence of $\mathbb{P}^n_{\theta^*}$ to $\mathbb{P}^m_{\theta^*}$ in terms of the Sinkhorn loss than MC, whereas for $d = 5$ (11b), there is no difference visible between the methods. Again, the curse of dimensionality is apparent when comparing the MC approach for $d = 1$ and $d = 5$, and the dimension-dependent constant in the QMC and RQMC error convergence rates seems to compound this effect for higher dimensions. Thus, we can conclude that the convergence of the Sinkhorn loss based estimator suffers from a much greater dimension dependence when using QMC and RQMC methods than the squared MMD.

5.2 Beta Distribution

The next example to be examined is the *beta distribution*. The beta distribution is a family of continuous probability distributions which are defined on the bounded interval $[0, 1]$ implying $d = 1$. The specific shape is controlled by the two parameters $\alpha > 0$ and $\beta > 0$ and the probability density function is given by

$$f(u) = \frac{u^{\alpha-1}(1-u)^{\beta-1}}{B(\alpha, \beta)}$$

where $u \in [0, 1]$ and $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and $\Gamma(\cdot)$ is the Gamma function. The beta function B serves as a normalisation constant ensuring that the total probability is 1. In the MMD framework, we have $\theta = (\theta_1, \theta_2)$ where $\theta_1 = \alpha$ and $\theta_2 = \beta$, so that $p = 2$. To obtain one data point from this distribution, we can first sample $u_k \overset{\text{IID}}{\sim} U([0, 1])$ for

$k = 1, \dots, \theta_1 + \theta_2$ and then use the following generator $G_\theta : [0,1]^{\theta_1+\theta_2} \to [0,1]$:

$$G_\theta(u_i) = \frac{\sum_{i=1}^{\theta_1} \log(u_i)}{\sum_{j=1}^{\theta_1+\theta_2} \log(u_i)}.$$

The use of this generator can be justified by noting that if z_1 and z_2 are Gamma$(\theta_1, 1)$ and Gamma$(\theta_2, 1)$ random variables respectively, then $z_1/(z_1 + z_2)$ is a Beta(θ_1, θ_2) random variable. Moreover, if $u \sim U([0,1]^{\theta_1})$ then $-\sum_{i=1}^{\theta_1} \log(u_i)$ is a Gamma$(\theta_1, 1)$ random variable (see Devroye 2006, p.405). Figure 12 shows the distribution for a dataset of 5,000 points sampled using the described generator and the true parameters for the following experiments. We can write for this sample $\{y_j\}_{j=1}^m \sim \mathbb{Q}$ where $\mathbb{Q} = \mathbb{P}_{\theta^*}$ with $\theta^* = (2, 5)$ and $m = 5,000$. For the use with QMC and RQMC, the u_i are replaced by the corresponding QMC and RQMC point sets.

The beta distribution will be used to analyse whether the obtained results are consistent across varying QMC and RQMC point sets. In particular, we will compare the default choice for the experiments reported here, the Halton and scrambled Halton sequence, with the pairs Sobol' and digitally shifted Sobol' sequences, and rank-1 lattice and digitally shifted rank-1 lattice. For this purpose, we investigate the rate at which the empirical measure $\mathbb{P}_\theta^n$ converges to the true measure $\mathbb{P}_\theta$ in terms of their squared MMD. For the true measure, the approximation $\mathbb{P}_\theta^m$ with a large number of samples $m = 2^{14}$ is used. Similar to the previous experiments, $R = 25$ repetitions are averaged

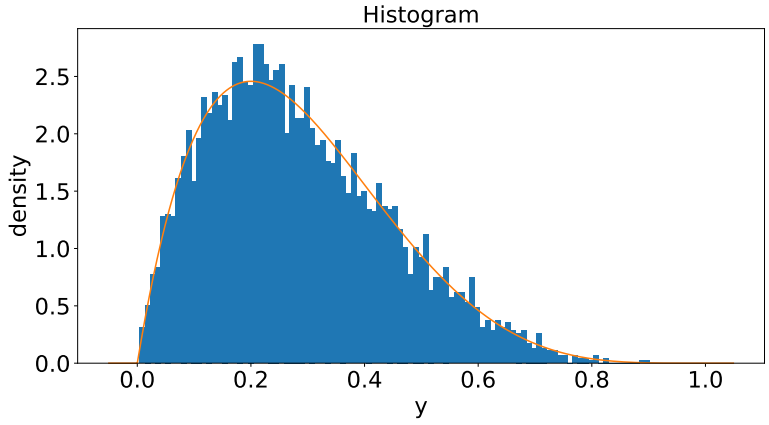
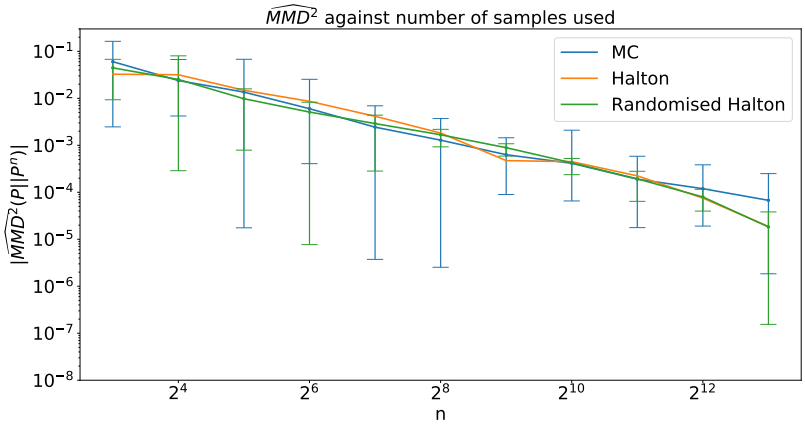
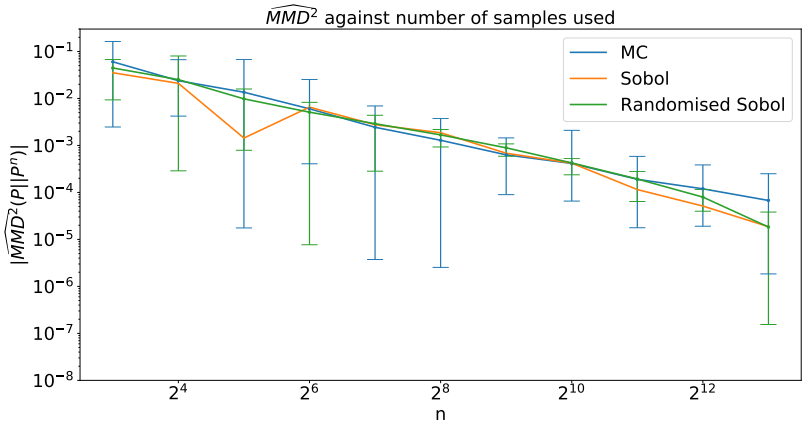


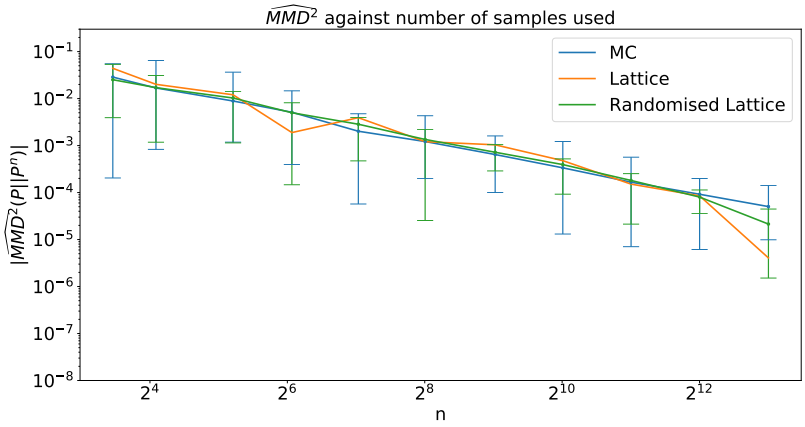
Figure 12: Histogram of 5,000 data points of the beta distribution with parameters $\theta_1 = 2$ and $\theta_2 = 5$ sampled using MC with the analytical density function in orange



(a) Halton (QMC) and scrambled Halton sequence (RQMC)



(b) Sobol (QMC) and digitally shifted Sobol sequence (RQMC)



(c) Rank-1 lattice (QMC) and shifted rank-1 lattice (RQMC)

Figure 13: Convergence of $\mathbb{P}_\theta^n$ to $\mathbb{P}_\theta^m$ in terms of their squared MMD comparing Halton, Sobol and Lattice point sets

for MC and RQMC per considered value for n and minimum and maximum values reported in the form of error bars. The considered sample sizes are powers of 2 for Halton and Sobol' sequences, while the closest prime is considered for the rank-1 lattice in order to accommodate the favourable sample sizes of the different construction approaches.

Figure 13 reports the results for Halton (13a), Sobol' (13b) and Lattice point sets (13c). In all settings, the MC approach achieves faster convergence for small n than QMC and RQMC. However, for sample sizes between 2^{11} and 2^{12}, the QMC and RQMC methods based on Halton, Sobol' and lattice point sets seem to approach the asymptotic QMC error convergence rate of n^{-1}, so that for larger n the convergence rate is faster than for MC. Considering the error bars, the results appear to be consistent across the examined low-discrepancy point sets, whereas the results for MC show considerable stochastic variation determining for which number of samples we can observe that QMC and RQMC methods becomes more efficient than MC for the one-dimensional beta distribution.

5.3 G-and-k Distribution

The next generative model we want to investigate is a synthetic example which is popular in literature: the univariate *g-and-k distribution* (see e.g. Fearnhead et al. 2012; Bernton et al. 2019). It is defined in terms of its quantile function $G_\theta(u) : [0,1)^2 \to \mathbb{R}$:

$$G_\theta(u) = a + b\left(1 + 0.8\frac{1-\exp(-gz(u))}{1+\exp(-gz(u))}\right)\left(1+z(u)^2\right)^k z(u)$$

where $z(u) = \sqrt{-2\log u_1}\cos(2\pi u_2)$ and $u_1, u_2 \overset{\text{IID}}{\sim} U([0,1))$, so that $z \sim N(0,1)$. The parameters a and $b > 0$ control location and scale of the distribution, while g and $k > -\frac{1}{2}$ measure skewness and kurtosis respectively allowing for a very flexible family of distributions (Rayner et al. 2002). Thus, the parameters of interest are $\theta_1 = a$, $\theta_2 = b$, $\theta_3 = g$ and, to avoid numerical instabilities, $\theta_4 = \exp(k)$ giving the parameter vector $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$ with $p = 4$. The partial derivatives of the generator $\nabla_\theta G_\theta$ required for the optimisation are given in appendix B.2. Figure 14 visualises the g-and-k distribution with $\theta^* = (3, 1, 1, -\log(2))$, which is also the true parameter in the following experiments. Further, we consider a dataset consisting of samples $\{y_j\}_{j=1}^m \overset{\text{IID}}{\sim} \mathbb{Q}$ where

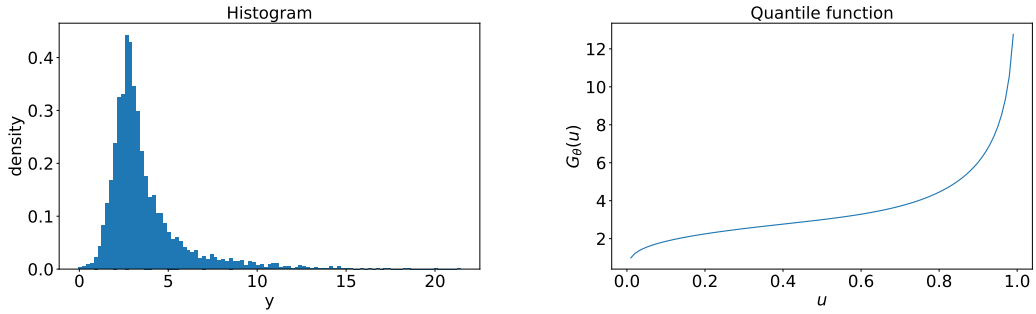


Figure 14: Histogram of 5,000 data points generated from the g-and-k distribution with $\theta = (3, 1, 1, -\log(2))$ (left) and the corresponding generator function (right)

$\mathbb{Q} = \mathbb{P}_{\theta^*}$ implying $d = 1$.

We begin the analysis of the g-and-k distribution in the minimum MMD framework by presenting the optimisation results for MC, QMC and RQMC when using SGD in a setting, in which θ^* is utilised to generate a large number of $m = 2^{11}$ samples from $\mathbb{Q}$. For all methods, the optimisation algorithm is run for 10,000 iterations with $n = 2^9$ samples generated at each iteration. The step size is fixed to 0.1 to ensure convergence and $\theta^{(0)} = (0.5, 0.5, 0.5, 0.5)$ is picked as the start value.

The MMD loss for MC, QMC and RQMC at every iteration is visualised in Figure 16. It reveals a quick convergence for all methods as expected for a low dimensional optimisation problem. The speed of convergence appears to be faster for QMC and RQMC over the first descent steps, but the converged absolute loss is higher than MC

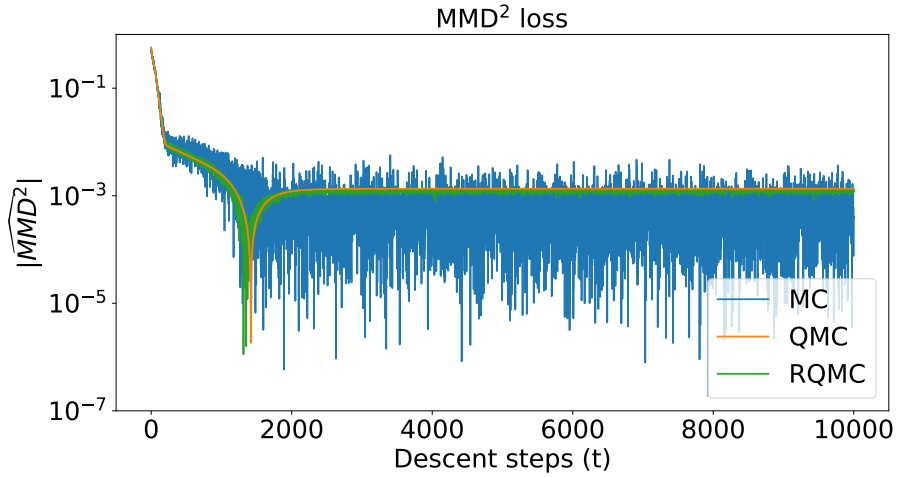


Figure 15: MMD loss when optimising using SGD

as the underlying squared MMD approximation is negative. To measure the precision of the different approaches more accurately, Figure 16 depicts the MSE for every estimated parameter against the number of descent steps for the MC, QMC and RQMC approaches. Compared to the results presented for the Gaussian location model, the parameters of the g-and-k distribution appear to be more difficult to estimate as only the MSE for θ_1 is converging quickly and monotonically to a value close to zero. However, the lines representing the MSE are widely overlapping for all converged estimates. Thus, in terms of precision, MC, QMC and RQMC approaches show similar performances for

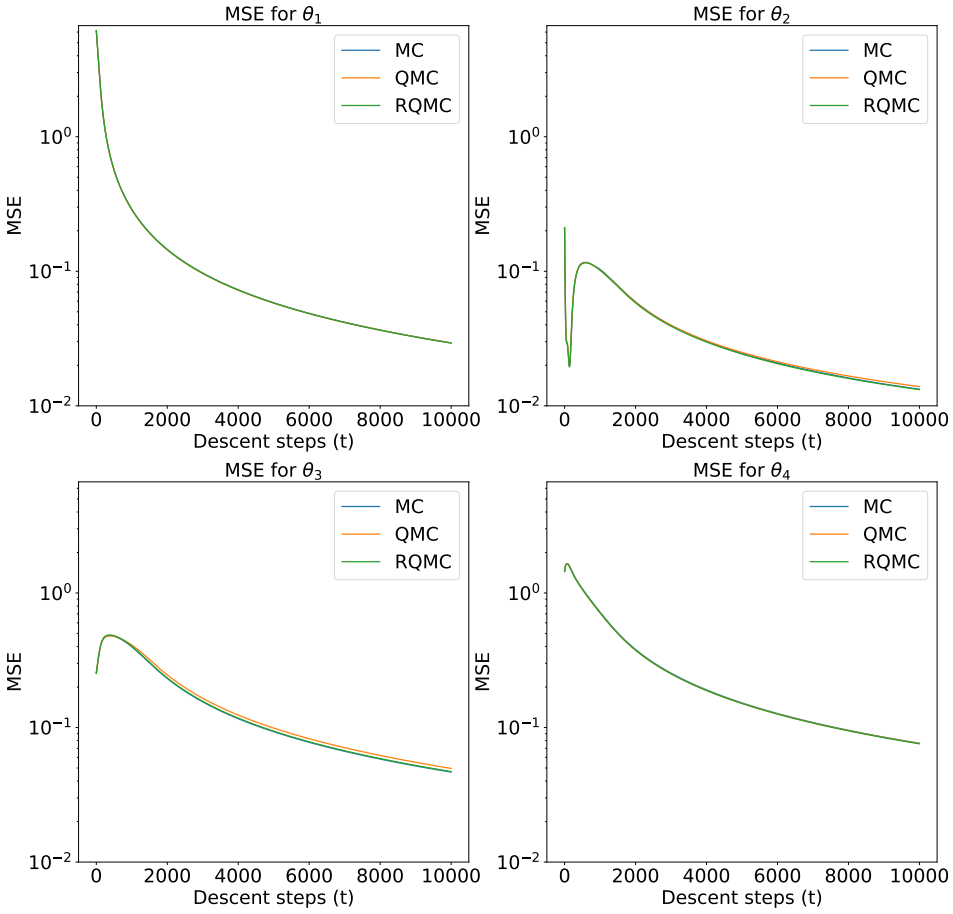


Figure 16: MSE when optimising using SGD

the g-and-k distribution when estimating the parameters using SGD optimisation.

Due to numerical instabilities when estimating θ_3 and θ_4 via NSGD, no optimisation results are reported using this alternative optimisation algorithm. This issue is also recognisable in the application of the g-and-k distribution by Briol et al. (2019, Figure 3), where the estimates obtained using NSGD meander around the true value but fail to converge.

For the g-and-k distribution, we also want to investigate whether we can notice differences between MC, QMC and RQMC in the convergence rate of the empirical measure $\mathbb{P}_\theta^n$ to the true measure $\mathbb{P}_\theta$ approximated by $\mathbb{P}_\theta^m$ in terms of their squared MMD. For this experiment, m is set to 2^{13} and $R = 25$ repetitions are averaged per considered sample size n for MC and QMC with maximum and minimum values reported as error bars. The results are provided in Figure 17. With QMC and RQMC approaches behaving similarly, the MC approach converges faster on average for $n < 2^{10}$, whereas for $n > 2^{11}$ QMC and RQMC exhibit faster convergence rates. The variation in the RQMC results decreases considerably in n, whereas the MC approach is subject to continuous variability. We can conclude that for large enough n the MMD-based estimator can again leverage the faster error convergence rates of QMC and RQMC in a one-dimensional setting.

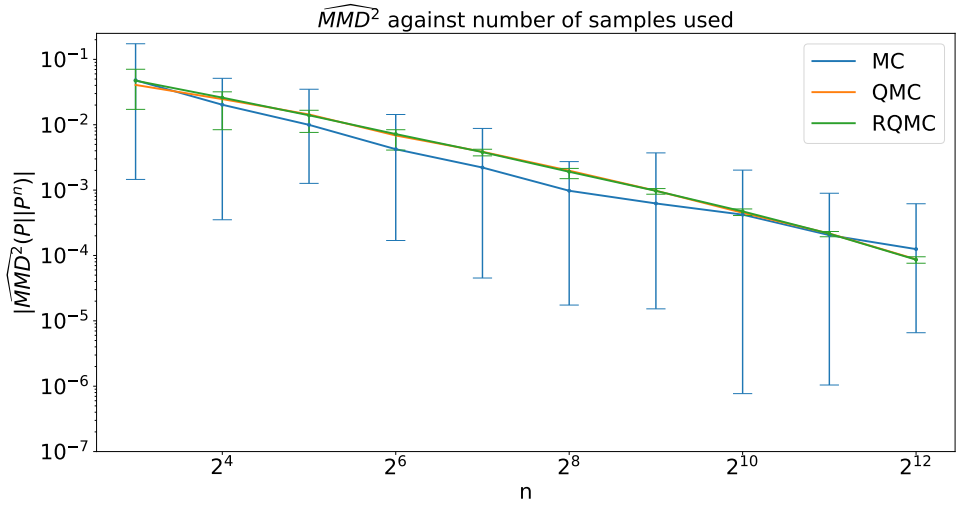


Figure 17: MMD^2 against a range of values for n

6 Discussion

The preceding numerical experiments demonstrated that the use of QMC and RQMC methods provides faster convergence of the squared MMD in low-dimensional settings for the Gaussian location model, beta distribution and g-and-k distribution for sample sizes larger than 2^{11} or 2^{12}. We will now discuss a number of possible explanations for why we could not observe this result in higher dimensions.

First, it is important to keep in mind that the convergence rates of QMC and RQMC are only asymptotic implying that we do not know for what sample size we could expect QMC and RQMC to converge faster than MC (Drew et al. 2006). Considering the convergence rate of the squared MMD, this means that the investigated sample sizes might simply be not large enough to capture the number of samples for which QMC and RQMC based approaches become more efficient than the MC one in higher dimensions.

This issue also has a strong link to the dimensionality of the considered generative model. Recalling the QMC convergence rate of $O(n^{-1}(\log n)^d)$, the dimension dependence of the constant term is obvious. Thus, for the asymptotic convergence rate to take effect for high-dimensional problems in the minimum MMD framework, the necessary sample size increases with the dimension. This could explain why we were only able to demonstrate the higher efficiency of the use of QMC and RQMC point sets for up to three dimensions. As a remedy for this dimension dependence, further extensions of QMC methods have been developed following the notion of 'effective dimension', which might be considered to improve the scalability of the estimator regarding the model complexity: weighted spaces have been introduced to accommodate situations in which some coordinate directions are more important than others in the sense of higher variability providing error bounds which are independent of d (Sloan et al. 1998). An example are the so-called weighted Sobolev spaces which refer to weighted spaces of non-periodic functions (Dick et al. 2013). If the weights decrease sufficiently rapidly, a convergence rate arbitrarily close to $O(n^{-1})$ can be attained with the implied constant being independent of dimension (Sloan et al. 1998). Based on this idea, several of these higher order QMC rules are available for sufficiently smooth functions. See for example Goda et al. (2020) for a recent review. Alternatively, the approach of Drew et al. (2006) could be investigated further: they suggest to use a combination of QMC and MC point sets

to avoid the outlined dimensionality problem, while enhancing the performance by using the QMC points for the most important dimensions of the integrand.

At the same time, there are also practical issues leading to the limitations of the conducted experiments. When studying the convergence of the squared MMD, we cannot eliminate the possibility that the approximation of the true measure might not be precise enough to demonstrate the desired results. A larger number of samples m might be needed to attain the required precision. However, due to limited computational resources this was not feasible. For the same reason, it was not possible to analyse the convergence of the squared MMD for larger n.

Generally, the settings for the numerical experiments were kept fixed in terms of kernel choice and the corresponding lengthscale to isolate the effects of using QMC and RQMC point sets. However, this means that no conclusions with respect to whether the observed results also carry over into other settings can be drawn. Similarly, it was not explored whether the tuning of optimisation parameters could improve the performance of SGD and NSGD algorithms and whether effects vary between the use of MC, QMC and RQMC methods. Enhancements to consider could encompass employing an adaptive step size or adding momentum (Bottou et al. 2018).

A practical issue which was neglected in the numerical experiments is the impact of switching from MC to QMC and RQMC methods on the computational cost. This is especially relevant for large and complex generative models, which require large numbers of samples for precise inference. Further investigations should therefore take this problem into consideration as is has important implications for the applicability of QMC and RQMC in the minimum MMD framework.

7 Conclusion

Modern computational statistics requires inference tools to scale well with model complexity and data size. In the case of intractable generative models, classical approaches to inference such as ML estimation are not readily applicable, so that likelihood-free methods such as minimum divergence estimators are needed. The goal of this thesis was to explore the possibilities provided by QMC methods to further improve the scalability

of a specific instance of minimum divergence estimators: the minimum MMD estimators proposed by Briol et al. (2019).

To do so, basic concepts around RKHS and distances between probability distributions were reviewed to introduce the general framework of minimum divergence estimators with a special focus on generative models. Equipped with this background, two different approaches for statistical inference for generative models were presented. In detail, the minimum MMD estimators of Briol et al. (2019) were covered including the presentation of two alternative optimisation algorithms and statistical properties. Then, the Sinkhorn loss of Genevay et al. (2018) was briefly introduced as a main competitor. Motivated by the necessity to approximate the arising integrals, an introduction to MC, QMC and RQMC methods for integration was given. Combining the previously reviewed concepts, a novel application of QMC methods in the minimum MMD framework of Briol et al. (2019) was suggested to improve the efficiency of the estimators. Finally, the results of a range of numerical experiments for three different generative models were presented and implications as well as limitations discussed.

In the numerical experiments, QMC and RQMC were found to be successfully applicable for the optimisation of the minimum MMD estimator using both SGD and NSGD algorithms and work at least as well as MC in terms of precision for the considered generative models. For the Gaussian location model with up to three dimensions and the one-dimensional beta and g-and-k distributions, it was shown that for a large enough number of samples, the squared MMD convergence rate using QMC and RQMC approaches is faster than for MC. Thus, the modified estimators appear to scale better with data size than the MC-based version. Moreover, a strong negative impact of expanding the dimension on the convergence rate of the squared MMD when using QMC and RQMC was observed which can possibly be explained by QMC and RQMC error convergence rates only holding asymptotically. This implies that further research should target improving the scalability of the modified estimators in terms of model complexity. Furthermore, consistent results for different QMC and RQMC point sets were reported. Considering Wasserstein distance based estimators, the convergence rate of the 1-Wasserstein distance and the Sinkhorn loss was found to speed up considerably using QMC and RQMC point sets for low-dimensional Gaussian location models. For

higher dimensions, the curse of dimensionality for these distances was compounded by the asymptotic nature of QMC and RQMC error convergence rates erasing the efficiency gains by QMC and RQMC for small numbers of samples which were observed for low dimensions.

Future work should not only focus on a theoretical foundation for the observed results, but should also address the seemingly strong dimension-dependence of benefits from using QMC and RQMC methods with minimum MMD estimators. A promising route could be the employment of higher order QMC methods providing error convergence rates that are independent of dimension (Goda et al. 2020).

Moreover, the numerical experiments also touched upon the advantages of using QMC and RQMC methods to increase the efficiency of inference for generative models using the Sinkhorn loss of Genevay et al. (2018). However, the conducted analysis only considered a very small value for the regularisation parameter ϵ, for which the Sinkhorn divergence is very close to the Wasserstein distance. Since for increasing ϵ, the Sinkhorn divergence approaches the MMD (Genevay et al. 2019) and is thus less susceptible to the curse of dimensionality, future work could investigate how QMC and RQMC methods impact the convergence rate of the Sinkhorn loss for larger ϵ.

References

Adams, Malcolm and Victor Guillemin (1996). *Measure theory and probability*. Reprinted with corrections from the 1986 Wadsworth edition. Boston: Birkhäuser.

Aistleitner, Christoph, Markus Hofer and Volker Ziegler (2014). 'On the uniform distribution modulo 1 of multidimensional LS-sequences'. In: *Annali di Matematica Pura ed Applicata (1923 -)* 193.5, pp. 1329–1344.

Amari, Shun-ichi (1998). 'Natural gradient works efficiently in learning'. In: *Neural Computation* 10.2, pp. 251–276.

Amari, Shun-ichi (2016). *Information geometry and its applications*. 1st ed. 2016. Vol. 194. Applied Mathematical Sciences. Tokyo: Springer.

Arjovsky, Martin, Soumith Chintala and Léon Bottou (2017). 'Wasserstein generative adversarial networks'. In: *International Conference on Machine Learning*, pp. 214–223.

Bassetti, Federico, Antonella Bodini and Eugenio Regazzini (2006). 'On minimum Kantorovich distance estimators'. In: *Statistics & Probability Letters* 76.12, pp. 1298–1302.

Basu, Ayanendranath, Ian R. Harris and Srabashi Basu (1997). '2 Minimum distance estimation: The approach using density-based distances'. In: *Robust Inference*. Ed. by Gangadharrao Soundalyarao Maddala and Calyampudi Radhakrishna Rao. Vol. 15. Handbook of Statistics. Elsevier, pp. 21–48.

Basu, Ayanendranath, Ian R. Harris, Nils L. Hjort and M. C. Jones (1998). 'Robust and efficient estimation by minimising a density power divergence'. In: *Biometrika* 85.3, pp. 549–559.

Basu, Ayanendranath, Chanseok Park and Hiroyuki Shioya (2011). *Statistical inference: The minimum distance approach*. Vol. 120. Monographs on statistics and applied probability. Boca Raton, Fla: Chapman & Hall/CRC.

Berlinet, Alain and Christine Thomas-Agnan (2004). *Reproducing kernel Hilbert spaces in probability and statistics*. New York: Springer Science+Business Media.

Bernardo, José M. and Adrian F. M. Smith (1994). *Bayesian theory*. Wiley series in probability and mathematical statistics. Chichester: Wiley.

Bernton, Espen, Pierre E. Jacob, Mathieu Gerber and Christian P. Robert (2019). 'Approximate Bayesian computation with the Wasserstein distance'. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 81.2, pp. 235–269.

Besag, Julian (1974). 'Spatial interaction and the statistical analysis of lattice systems'. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2, pp. 192–225.

Borgwardt, Karsten M., Arthur Gretton, Malte J. Rasch, Hans-Peter Kriegel, Bernhard Schölkopf and Alexander J. Smola (2006). 'Integrating structured biological data by kernel maximum mean discrepancy'. In: *Bioinformatics (Oxford, England)* 22.14, e49–57.

Bottou, Léon, Frank E. Curtis and Jorge Nocedal (2018). 'Optimization methods for large-scale machine learning'. In: *SIAM Review* 60.2, pp. 223–311.

Box, George E. P. and Mervin E. Muller (1958). 'A note on the generation of random normal deviates'. In: *The Annals of Mathematical Statistics* 29.2, pp. 610–611.

Boyle, Phelim P., Yongzeng Lai and Ken Seng Tan (2005). 'Pricing options using lattice rules'. In: *North American Actuarial Journal* 9.3, pp. 50–76.

Briol, François-Xavier, Alessandro Barp, Andrew B. Duncan and Mark Girolami (2019). 'Statistical inference for generative models with maximum mean discrepancy'. In: *arXiv preprint arXiv:1906.05944*.

Caflisch, Russel E. (1998). 'Monte Carlo and quasi-Monte Carlo methods'. In: *Acta Numerica* 7, pp. 1–49.

Cambou, Mathieu, Marius Hofert and Christiane Lemieux (2017). 'Quasi-random numbers for copula models'. In: *Statistics and Computing* 27.5, pp. 1307–1329.

Choi, Sou-Cheng T., Fred J. Hickernell, Yuhan Ding, Lan Jiang, Lluís Antoni Jiménez Rugama, Da Li, Jagadeeswaran Rathinavel, Xin Tong, Kan Zhang, Yizhi Zhang and Xuan Zhou (2020a). *GAIL: Guaranteed Automatic Integration Library, MATLAB Software*.

Choi, Sou-Cheng T., Fred J. Hickernell, M. McCourt and A. Sorokin (2020b). *QMCPy: A quasi-Monte Carlo Python Library*.

Cichocki, Andrzej and Shun-ichi Amari (2010). 'Families of Alpha- Beta- and Gamma-divergences: Flexible and robust measures of similarities'. In: *Entropy* 12.6, pp. 1532–1568.

Cohn, Donald L. (2013). *Measure Theory: Second Edition*. 2nd ed. 2013. Birkhäuser Advanced Texts Basler Lehrbücher. New York, NY and s.l.: Springer New York.

Cuturi, Marco (2013). 'Sinkhorn distances: Lightspeed computation of optimal transport'. In: *Advances in neural information processing systems*, pp. 2292–2300.

Dawid, A. Philip (2007). 'The geometry of proper scoring rules'. In: *Annals of the Institute of Statistical Mathematics* 59.1, pp. 77–93.

Devroye, Luc (2006). 'Chapter 4 Nonuniform random variate generation'. In: *Simulation*. Ed. by Shane G. Henderson and Barry L. Nelson. Vol. 13. Handbooks in Operations Research and Management Science. Elsevier, pp. 83–121.

Dick, Josef, Frances Y. Kuo and Ian H. Sloan (2013). 'High-dimensional integration: The quasi-Monte Carlo way'. In: *Acta Numerica* 22, pp. 133–288.

Dick, Josef and Friedrich Pillichshammer (2010). *Digital nets and sequences: Discrepancy and quasi-Monte Carlo integration*. Cambridge: Cambridge University Press.

Drew, Shane S. and Tito Homem-de-Mello (2006). 'Quasi-Monte Carlo strategies for stochastic optimization'. In: *Proceedings of the 2006 Winter Simulation Conference*, pp. 774–782.

Dudley, Richard M. (2002). *Real analysis and probability*. Second edition. Vol. 74. Cambridge studies in advanced mathematics. Cambridge: Cambridge University Press.

Dunham, William (2015). *The calculus gallery: Masterpieces from Newton to Lebesgue*. Princeton: Princeton University Press.

Dziugaite, Gintare Karolina, Daniel M. Roy and Zoubin Ghahramani (2015). 'Training generative neural networks via maximum mean discrepancy optimization'. In: *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*. UAI'15. Arlington, Virginia, USA: AUAI Press, pp. 258–267.

El Moselhy, Tarek A. and Youssef M. Marzouk (2012). 'Bayesian inference with optimal maps'. In: *Journal of Computational Physics* 231.23, pp. 7815–7850.

Faure, Henri (1951). 'Discrépances de suites associées à un système de numération (en dimension un)'. In: *Bulletin de la Société mathématique de France* 79, pp. 143–182.

Fearnhead, Paul and Dennis Prangle (2012). 'Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation'. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.3, pp. 419–474.

Fukumizu, Kenji, Arthur Gretton, Gert R. Lanckriet, Bernhard Schölkopf and Bharath K. Sriperumbudur (2009). 'Kernel choice and classifiability for RKHS embeddings of probability distributions'. In: *Advances in Neural Information Processing Systems 22*. Ed. by Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams and A. Culotta. Curran Associates, Inc, pp. 1750–1758.

Fukumizu, Kenji, Le Song and Arthur Gretton (2013). 'Kernel Bayes' rule: Bayesian inference with positive definite kernels'. In: *Journal of Machine Learning Research* 14.82, pp. 3753–3783.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari and Donald B. Rubin (2014). *Bayesian data analysis*. Third edition. Texts in statistical science series. Boca Raton, London and New York: CRC Press Taylor and Francis Group.

Genevay, Aude, Lenaic Chizat, Francis Bach, Marco Cuturi and Gabriel Peyré (2019). 'Sample complexity of Sinkhorn divergences'. In: *AISTATS'19 - 22nd International Conference on Artificial Intelligence and Statistics*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Okinawa, Japan.

Genevay, Aude, Gabriel Peyre and Marco Cuturi (2018). 'Learning generative models with Sinkhorn divergences'. In: *Proceedings of Machine Learning Research*. Ed. by Amos Storkey and Fernando Perez-Cruz. Vol. 84. Proceedings of Machine Learning Research. Playa Blanca, Lanzarote, Canary Islands: PMLR, pp. 1608–1617.

Giles, Mike, Frances Y. Kuo, Ian H. Sloan and Benjamin J. Waterhouse (2008). 'Quasi-Monte Carlo for finance applications'. In: *ANZIAM Journal* 50, p. 308.

Gneiting, Tilmann and Adrian E. Raftery (2007). 'Strictly proper scoring rules, prediction, and estimation'. In: *Journal of the American Statistical Association* 102.477, pp. 359–378.

Gnewuch, Michael, Anand Srivastav and Carola Winzen (2009). 'Finding optimal volume subintervals with k points and calculating the star discrepancy are NP-hard problems'. In: *Journal of Complexity* 25.2, pp. 115–127.

Goda, Takashi and Kosuke Suzuki (2020). '4. Recent advances in higher order quasi-Monte Carlo methods'. In: *Discrepancy Theory*. Ed. by Dmitriy Bilyk, Josef Dick and Friedrich Pillichshammer. De Gruyter, pp. 69–102.

Gretton, Arthur, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf and Alexander J. Smola (2012). 'A Kernel Two-Sample Test'. In: *Journal of Machine Learning Research* 13.25, pp. 723–773.

Haker, Steven, Lei Zhu, Allen Tannenbaum and Sigurd Angenent (2004). 'Optimal mass transport for registration and warping'. In: *International Journal of Computer Vision* 60.3, pp. 225–240.

Hampel, Frank R. (1971). 'A general qualitative definition of robustness'. In: *The Annals of Mathematical Statistics* 42.6, pp. 1887–1896.

Heiss, Florian and Viktor Winschel (2008). 'Likelihood approximation by numerical integration on sparse grids'. In: *Journal of Econometrics* 144.1, pp. 62–80.

Hickernell, Fred J. (1996). 'The mean square discrepancy of randomized nets'. In: *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 6.4, pp. 274–296.

Hofert, Marius and Christiane Lemieux (2016). *qrng: (Randomized) Quasi-Random Number Generators*.

Hofmann, Thomas, Bernhard Schölkopf and Alexander J. Smola (2008). 'Kernel methods in machine learning'. In: *The Annals of Statistics* 36.3, pp. 1171–1220.

Huber, Peter J. and Elvezio M. Ronchetti (2011). *Robust statistics*. 2nd Ed. Vol. 693. Wiley Series in Probability and Statistics. Wiley.

Jank, Wolfgang (2005). 'Quasi-Monte Carlo sampling to improve the efficiency of Monte Carlo EM'. In: *Computational Statistics & Data Analysis* 48.4, pp. 685–701.

Kantorovitch, Leonid (1958). 'On the translocation of masses'. In: *Management Science* 5.1, pp. 1–4.

Korobov, Nikolai M. (1959). 'The approximate computation of multiple integrals'. In: *Dokl. Akad. Nauk SSSR*. Vol. 124, pp. 1207–1210.

Kreyszig, Erwin (1989). *Introductory functional analysis with applications*. Wiley classics library. New York: Wiley.

Kuipers, Lauwerens and Harald Niederreiter (1974). *Uniform distribution of sequences*. Pure and applied mathematics. New York: Wiley.

Kuo, Frances Y. (2003). 'Component-by-component constructions achieve the optimal rate of convergence for multivariate integration in weighted Korobov and Sobolev spaces'. In: *Journal of Complexity* 19.3, pp. 301–320.

Kuo, Frances Y. and Ian H. Sloan (2005). 'Lifting the curse of dimensionality'. In: *Notices of the AMS* 52, pp. 1320–1329.

Kuo, Frances Y., Grzegorz W. Wasilkowski and Benjamin J. Waterhouse (2006). 'Randomly shifted lattice rules for unbounded integrands'. In: *Journal of Complexity* 22.5, pp. 630–651.

L'Ecuyer, Pierre (1994). 'Efficiency improvement and variance reduction'. In: *Proceedings of Winter Simulation Conference*, pp. 122–132.

L'Ecuyer, Pierre (2009). 'Quasi-Monte Carlo methods with applications in finance'. In: *Finance and Stochastics* 13.3, pp. 307–349.

L'Ecuyer, Pierre, Christian Lécot and Bruno Tuffin (2008). 'A randomized quasi-Monte Carlo simulation method for Markov Chains'. In: *Operations Research* 56.4, pp. 958–975.

Lemieux, Christiane (2009). *Monte Carlo and quasi-Monte Carlo sampling*. Springer Series in Statistics. New York, NY: Springer.

Li, Peihua, Qilong Wang and Lei Zhang (2013). 'A novel earth mover's distance methodology for image matching with Gaussian mixture models'. In: *2013 IEEE International Conference on Computer Vision*, pp. 1689–1696.

Li, Yujia, Kevin Swersky and Richard Zemel (2015). 'Generative moment matching networks'. In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. ICML'15. JMLR.org, pp. 1718–1727.

Marsaglia, George and Thomas A. Bray (1964). 'A convenient method for generating normal variables'. In: *SIAM Review* 6.3, pp. 260–264.

Mohamed, Shakir and Balaji Lakshminarayanan (2016). 'Learning in implicit generative models'. In: *arXiv preprint arXiv:1610.03483*.

Møller, Jesper, Anthony N. Pettitt, Robert Reeves and Kasper K. Berthelsen (2006). 'An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants'. In: *Biometrika* 93.2, pp. 451–458.

Monge, Gaspard (1781). 'Mémoire sur la théorie des déblais et des remblais'. In: *Histoire de l'Académie Royale des Sciences de Paris*.

Morokoff, William J. and Russel E. Caflisch (1995). 'Quasi-Monte Carlo integration'. In: *Journal of Computational Physics* 122.2, pp. 218–230.

Muandet, Krikamol, Kenji Fukumizu, Bharath Sriperumbudur, Arthur Gretton and Bernhard Schölkopf (2014). 'Kernel mean estimation and stein effect'. In: *International Conference on Machine Learning*, pp. 10–18.

Muandet, Krikamol, Kenji Fukumizu, Bharath Sriperumbudur and Bernhard Schölkopf (2017). 'Kernel mean embedding of distributions: A review and beyond'. In: *Foundations and Trends® in Machine Learning* 10.1-2, pp. 1–141.

Müller, Alfred (1997). 'Integral probability metrics and their generating classes of functions'. In: *Advances in Applied Probability* 29.2, pp. 429–443.

Murphy, Susan A. and Aad W. van der Vaart (2000). 'On profile likelihood'. In: *Journal of the American Statistical Association* 95.450, pp. 449–465.

Niederreiter, Harald (1987). 'Point sets and sequences with small discrepancy'. In: *Monatshefte für Mathematik* 104.4, pp. 273–337.

Niederreiter, Harald (1992). *Random number generation and quasi-Monte Carlo methods*. Society for Industrial and Applied Mathematics.

Norden, R. H. (1973). 'A survey of maximum likelihood estimation: Part 2'. In: *International Statistical Review / Revue Internationale de Statistique* 41.1, p. 39.

Ökten, Giray and Ahmet Göncü (2011). 'Generating low-discrepancy sequences from the normal distribution: Box–Muller or inverse transform?' In: *Mathematical and Computer Modelling* 53.5-6, pp. 1268–1281.

Owen, Art B. (1997). 'Monte Carlo variance of scrambled net quadrature'. In: *SIAM Journal on Numerical Analysis* 34.5, pp. 1884–1910.

Owen, Art B. (2005). 'Multidimensional variation for quasi-Monte Carlo'. In: *Contemporary Multivariate Analysis and Design of Experiments*. Ed. by Jianqing Fan and Gang Li. WORLD SCIENTIFIC, pp. 49–74.

Owen, Art B. (2020). 'On dropping the first Sobol' point'. In: *arXiv preprint arXiv:2008.08051*.

Pan, Jianxin and Robin Thompson (2007). 'Quasi-Monte Carlo estimation in generalized linear mixed models'. In: *Computational Statistics & Data Analysis* 51.12, pp. 5765–5775.

Park, Mijung, Wittawat Jitkrittum and Dino Sejdinovic (2016). 'K2-ABC: Approximate Bayesian computation with kernel embeddings'. In: *Proceedings of Machine Learning Research*. Ed. by Arthur Gretton and Christian C. Robert. Vol. 51. Proceedings of Machine Learning Research. Cadiz, Spain: PMLR, pp. 398–407.

Pele, Ofir and Michael Werman (2009). 'Fast and robust earth mover's distances'. In: *2009 IEEE 12th International Conference on Computer Vision*, pp. 460–467.

Peyré, Gabriel and Marco Cuturi (2019). 'Computational optimal transport: With applications to data science'. In: *Foundations and Trends® in Machine Learning* 11.5-6, pp. 355–607.

Pillards, Tim and Ronald Cools (2006). 'Using Box-Muller with low discrepancy points'. In: *Computational Science and Its Applications - ICCSA 2006*. Ed. by Marina L. Gavrilova, Osvaldo Gervasi, Vipin Kumar, C. J. Kenneth Tan, David Taniar, Antonio Laganá, Youngsong Mun and Hyunseung Choo. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 780–788.

Ramdas, Aaditya, Nicolás García Trillos and Marco Cuturi (2017). 'On Wasserstein two-sample testing and related families of nonparametric tests'. In: *Entropy* 19, p. 47.

Ramstead, Maxwell J. D., Axel Constant, Paul B. Badcock and Karl J. Friston (2019). 'Variational ecology and the physics of sentient systems'. In: *Physics of life reviews* 31, pp. 188–205.

Ratmann, Oliver, Ole Jørgensen, Trevor Hinkley, Michael Stumpf, Sylvia Richardson and Carsten Wiuf (2007). 'Using likelihood-free inference to compare evolutionary dynamics of the protein networks of H. pylori and P. falciparum'. In: *PLoS computational biology* 3.11, e230.

Rayner, Glen D. and Helen L. MacGillivray (2002). 'Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions'. In: *Statistics and Computing* 12.1, pp. 57–75.

Riesselman, Adam J., John B. Ingraham and Debora S. Marks (2018). 'Deep generative models of genetic variation capture the effects of mutations'. In: *Nature methods* 15.10, pp. 816–822.

Rubner, Yossi, Carlo Tomasi and Leonidas J. Guibas (2000). 'The earth mover's distance as a metric for image retrieval'. In: *International Journal of Computer Vision* 40.2, pp. 99–121.

Schölkopf, Bernhard and Alexander J. Smola (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning. Cambridge, Mass: MIT Press.

Serfling, Robert J. (1980). *Approximation theorems of mathematical statistics*. Wiley Series in Probability and Statistics. New York, NY: Wiley.

Sloan, Ian H. and Stephen Joe (1994). *Lattice methods for multiple integration*. Oxford: Clarendon Press.

Sloan, Ian H., Frances Y. Kuo and Stephen Joe (2003). 'Constructing Randomly Shifted Lattice Rules in Weighted Sobolev Spaces'. In: *SIAM Journal on Numerical Analysis* 40.5, pp. 1650–1665.

Sloan, Ian H. and Henryk Woźniakowski (1998). 'When are quasi-Monte Carlo algorithms efficient for high dimensional integrals?' In: *Journal of Complexity* 14.1, pp. 1–33.

Sobol', Il'ya Meerovich (1967). 'On the distribution of points in a cube and the approximate evaluation of integrals'. In: *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki* 7.4, pp. 784–802.

Sriperumbudur, Bharath K., Kenji Fukumizu and Gert R. G. Lanckriet (2011). 'Universality, characteristic kernels and RKHS embedding of measures'. In: *Journal of Machine Learning Research* 12.70, pp. 2389–2410.

Sriperumbudur, Bharath K., Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf and Gert R. G. Lanckriet (2010). 'Hilbert space embeddings and metrics on probability measures'. In: *The Journal of Machine Learning Research* 11, pp. 1517–1561.

Steinwart, Ingo and Johanna F. Ziegel (2019). 'Strictly proper kernel scores and characteristic kernels on compact spaces'. In: *Applied and Computational Harmonic Analysis*.

Tao, Terence (2011). *An introduction to measure theory*. Vol. 126. Graduate studies in mathematics. Providence, RI: American Math. Soc.

Varin, Cristiano, Nancy Reid and David Firth (2011). 'An overview of composite likelihood methods'. In: *Statistica Sinica* 21.1, pp. 5–42.

Villani, Cédric (2003). *Topics in optimal transportation*. Vol. 58. Graduate studies in mathematics. Providence, RI: American Math. Soc.

Weed, Jonathan and Francis Bach (2019). 'Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance'. In: *Bernoulli* 25.4A, pp. 2620–2648.

Williams, David (1991). *Probability with martingales*. Cambridge: Cambridge University Press.

A Introduction to Measure and Probability Theory

In order to be able to formally present the problems of MC and QMC integration, we briefly present the concepts of Riemann and Lebesgue integrals as well as the necessary background on measure and probability theory.

To introduce the concept of *Riemann integrable* functions we follow (Tao 2011, chapter 1). A Riemann integral approximates the area under the curve using vertical rectangles. We consider a real-valued function f defined on an interval [a,b], which is partitioned using sub-intervals $[x_i, x_{i+1}]$ where $a = x_0 < x_1 < \cdots < x_n = b$ and a corresponding sequence of numbers $t_i \in [x_i, x_{i+1}]$ for $i = 1, \ldots, n$. Defining $\Delta x_i = x_i - x_{i-1}$, the *Riemann sum* is given by $\sum_{i=1}^{n} f(t_i)\Delta x_i$, for which each term represents a rectangle with height $f(t_i)$ and width Δx_i. The *Riemann integral* is then the limit of these Riemann sums as the partitions get finer, that is $\max_i \Delta x_i \to 0$. If this limit exists, the function f is called *Riemann integrable* over $[a, b]$ with corresponding Riemann integral $\int_a^b f(x)\mathrm{d}x$. Riemann integrability on $[a, b]$ requires the function f to be bounded on this interval and continuous almost everywhere, that is, the Lebesque measure of the set of all points at which f is not continuous in $[a, b]$ is zero.

The *Lebesgue integral* can be considered as a generalisation of the Riemann integral, which partitions the range of the integrand into layers that are not necessarily rectangles thereby allowing the integration of a larger class of functions including unbounded integrands (Dunham 2015, pp.212-218). Before presenting the Lebesgue integral formally, we give a brief introduction into the necessary background in measure and probability theory.

We begin by defining the *σ-algebra* $\mathcal{F}$ on the non-empty set $\mathcal{X}$. The σ-algebra $\mathcal{F}$ is a non-empty collection of subsets of $\mathcal{X}$, for which $\mathcal{X} \in \mathcal{F}$ and which is closed under complements, i.e. $A \in \mathcal{F}$ implies $A^c \in \mathcal{F}$, as well as closed under countable unions, i.e. if $A_i \in \mathcal{F}$ for $i \in \mathbb{N}$, then also $\bigcup_{i\in\mathbb{N}} A_i \in \mathcal{F}$ (Williams 1991, pp.15-16). The pair $(\mathcal{X}, \mathcal{F})$ is now called a *measurable space* (Dudley 2002, p.251). A map $\mu : \mathcal{F} \to [0, \infty]$ is a *measure* on $(\mathcal{X}, \mathcal{F})$ if μ satisfies $\mu(\emptyset) = 0$, and $\mu\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \sum_{i\in\mathbb{N}} \mu(A_i)$ for $\{A_i\}_i \in \mathbb{N} \in \mathcal{F}$ being disjoint sets with $A = \bigcup_{i\in\mathbb{N}} A_i$ (Williams 1991, p.18). The triplet $(\mathcal{X}, \mathcal{F}, \mu)$ is referred to as a *measure space* (Dudley 2002, p.87). Further, the measure μ is called a *probability measure* if $\mu(\mathcal{X}) = 1$, and then the triplet $(\mathcal{X}, \mathcal{F}, \mu)$ becomes a *probability*

space (Dudley 2002, p.251). Given two measurable spaces $(\mathcal{X}_1, \mathcal{F}_1)$ and $(\mathcal{X}_2, \mathcal{F}_2)$, a function $h : \mathcal{X}_1 \to \mathcal{X}_2$ with $h^{-1}(A) := \{x \in \mathcal{X}_1 : h(x) \in A\}$ for $A \subseteq \mathcal{X}_2$ is called $\mathcal{F}_1/\mathcal{F}_2$-*measurable* if $h^{-1} : \mathcal{F}_2 \to \mathcal{F}_1$, i.e. $h^{-1}(A) \in \mathcal{F}_1$ for all $A \in \mathcal{F}_2$ (Williams 1991, p.29-31). Is the σ-algebra with respect to which a function is measurable clear from the context, the function h is usually simply referred to as being *measurable*.

The *Lebesgue integral* of a measurable function f with respect to the Lebesgue measure μ over the measurable space $\mathcal{X}$ can be written as $\int_{\mathcal{X}} f \mathrm{d}\mu$. It can be constructed from *simple functions* which are finite linear combinations of indicator functions that can be used to approximate a measurable function from below. The presentation of this construction follows Adams et al. (1996, chapter 2). For $a_i \in [0, \infty]$ and $A_i \in \mathcal{F}$, a measurable simple function is defined as $\tilde{f}(x) = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i}$ where $\mathbb{1}_{A_i}$ denotes an indicator function with value 1 if $x \in A_i$ and 0 otherwise, so that $\int \mathbb{1}_A \mathrm{d}\mu = \mu(A)$. Then, the integral of a simple function is given by $\int \tilde{f} \mathrm{d}\mu = \sum_{i=1}^{n} a_i \mu(A_i)$. A positive measurable function can now be defined as a supremum of approximations based on simple functions, since if a positive measurable sequence $\{f_n\}_{n \in \mathbb{N}}$ converges pointwise to f, $\int f_n \mathrm{d}\mu \to \int f \mathrm{d}\mu$ holds (monotone convergence theorem). A not necessarily positive measurable function f can be expressed as the difference of two integrals of positive measurable functions by defining $f_+ = \max(f(x), 0)$ and $f_- = \max(-f(x), 0)$, so that we can write $\int f \mathrm{d}\mu = \int f_+ \mathrm{d}\mu - \int f_- \mathrm{d}\mu$ if at least one of $\int f^+ \mathrm{d}\mu$ and $\int f^- \mathrm{d}\mu$ is finite. If $\int |f| \mathrm{d}\mu < \infty$ where $|f| = f_+ + f_-$, f is considered *Lebesgue integrable*.

B Additional Material for Numerical Experiments

B.1 Gaussian Location Model

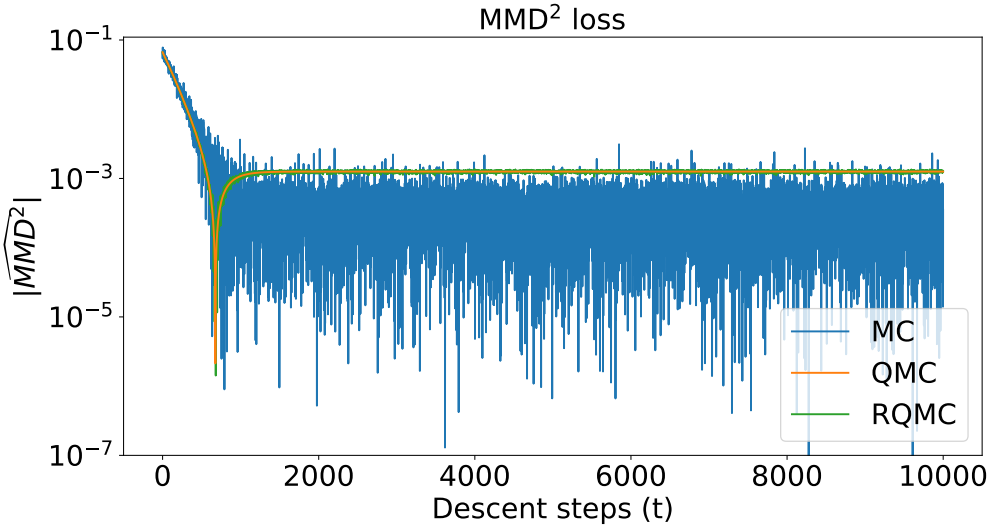


Figure 18: MMD loss when optimising using SGD for $d = 5$

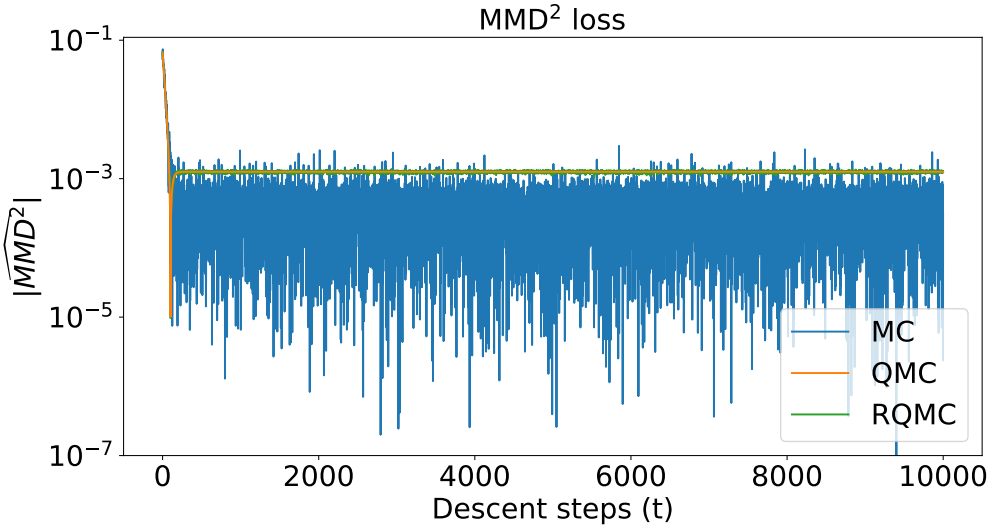


Figure 19: MMD loss when optimising using NSGD for $d = 5$

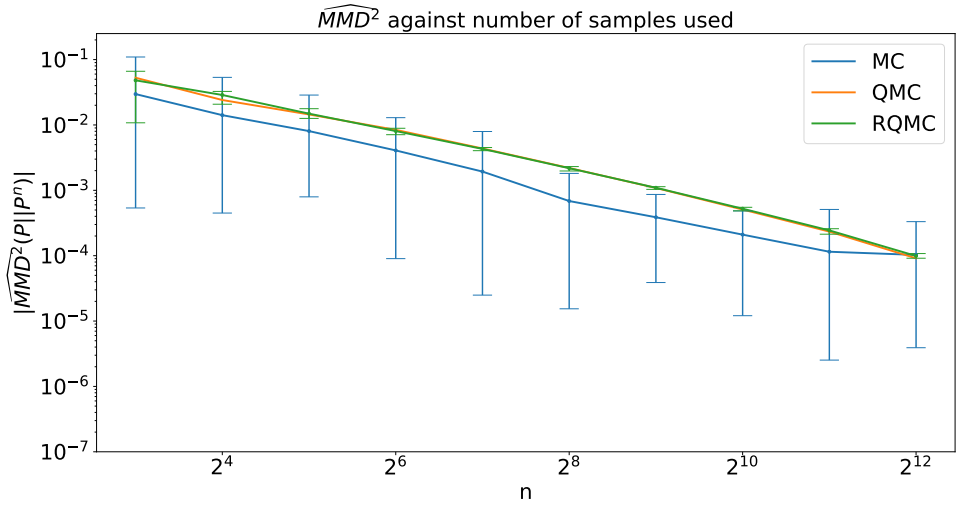


Figure 20: Convergence of $\mathbb{P}^n_{\boldsymbol{\theta}^*}$ to $\mathbb{P}^m_{\boldsymbol{\theta}^*}$ in terms of their squared MMD with $m = 2^{13}$ and $d = 3$

B.2 G-and-k distribution

For the optimisation procedure of the g-and-k distribution, the partial derivatives of the generator $\nabla_\theta G_\theta$ need to be accessible. Briol et al. (2019) provides those in appendix D.2:

$$\frac{\partial G_\theta(u)}{\partial \theta_1} = 1$$

$$\frac{\partial G_\theta(u)}{\partial \theta_2} = \left(1 + \frac{4\big(1 - \exp(-\theta_3 z(u))\big)}{5\big(1 - \exp(-\theta_3 z(u))\big)}\right)\big(1 + z(u)^2\big)^{\theta_4} z(u)$$

$$\frac{\partial G_\theta(u)}{\partial \theta_3} = \frac{8}{5}\theta_2 \frac{\exp(\theta_3 z(u))}{\big(1 + \exp(\theta_3 z(u))\big)^2}\big(1 + z(u)^2\big)^{\theta_4} z(u)^2$$

$$\frac{\partial G_\theta(u)}{\partial \theta_4} = \theta_2\left(1 + 0.8\frac{\big(1 - \exp(-\theta_3 z(u))\big)}{\big(1 + \exp(-\theta_3 z)\big)}\right)\big(1 + z(u)^2\big)^{\theta_4} \log(1 + z(u)^2) z(u)$$