Optimally-weighted Estimators of the Maximum Mean Discrepancy for Likelihood-Free Inference

Ayush Bharti [1] **Masha Naslidnyk** [2] **Oscar Key** [2] **Samuel Kaski** [1 3] **François-Xavier Briol** [2]

Abstract

Likelihood-free inference methods typically make use of a distance between simulated and real data. A common example is the maximum mean discrepancy (MMD), which has previously been used for approximate Bayesian computation, minimum distance estimation, generalised Bayesian inference, and within the nonparametric learning framework. The MMD is commonly estimated at a root-m rate, where m is the number of simulated samples. This can lead to significant computational challenges since a large m is required to obtain an accurate estimate, which is crucial for parameter estimation. In this paper, we propose a novel estimator for the MMD with significantly improved sample complexity. The estimator is particularly well suited for computationally expensive smooth simulators with low- to mid-dimensional inputs. This claim is supported through both theoretical results and an extensive simulation study on benchmark simulators.

1. Introduction

Many domains of science, medicine and engineering use our mechanistic understanding of real-world phenomena to create simulators that can represent system behaviour in different circumstances. Such simulator-based models define a stochastic procedure that can generate (possibly complex) synthetic data-sets, and are widely used in fields such as population genetics (Beaumont, 2010), ecology (Wood, 2010), astronomy (Cameron & Pettitt, 2012; Akeret et al., 2015), epidemiology (Kypraios et al., 2017), atmospheric contamination (Kopka et al., 2016), radio propagation (Bharti et al., 2022a), and agent-based modelling (Jennings, 1999). However, the ease of simulating data from the model comes at the cost of an intractable likelihood function, rendering most standard statistical inference methods inapplicable to such models. To solve this issue, a host of *likelihood-free inference* methods have been developed that circumvent the need to evaluate the likelihood or its derivatives, see Lintusaari et al. (2017); Cranmer et al. (2020) for an overview.

[1]Department of Computer Science, Aalto University, Espoo, Finland [2]Department of Statistical Science, University College London, London, United Kingdom [3]Department of Computer Science, University of Manchester, Manchester, United Kingdom. Correspondence to: Ayush Bharti <ayush.bharti@aalto.fi>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

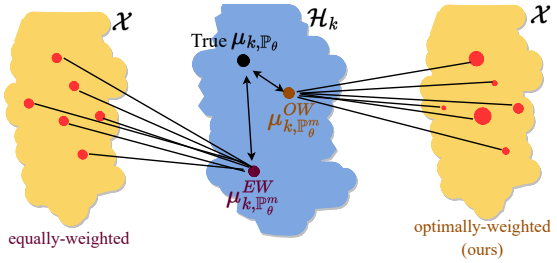


Figure 1. Estimating the MMD requires approximating the embedding $\mu_{k,\mathbb{P}_\theta}$ of the model $\mathbb{P}_\theta$ in a reproducing kernel Hilbert space $\mathcal{H}_k$. The classical approach consists of doing this from m equally-weighted independent samples from $\mathbb{P}_\theta$ (denoted $\mu^{EW}_{k,\mathbb{P}^m_\theta}$), but we show in this paper that it is possible to improve this estimator by using optimally-weighted samples (denoted $\mu^{OW}_{k,\mathbb{P}^m_\theta}$).

A common approach for likelihood-free inference involves comparing simulated observations from the model and the observed data, with respect to some notion of distance. Accurately estimating the distance is essential for inference but doing so usually requires simulating large amounts of synthetic data. This can be a computational bottleneck, especially for expensive simulators, which in the most extreme cases can take up to hundreds or thousands of CPU hours per simulation; see Niederer et al. (2019) for an example in cardiac modelling. Other examples include tsunami models based on shallow water equations that require several GPU hours per run (Behrens & Dias, 2015), runaway electron analysis models for nuclear fusion devices that require 24 CPU hours per run (Hoppe et al., 2021), and models of large-scale wind farms that require 100 CPU hours per run (Kirby et al., 2022). Naturally, the discrepancies popular for likelihood-free inference are those which can be efficiently estimated given samples from two distributions, such as the KL divergence (Jiang, 2018), Wasserstein dis-

tance (Peyré & Cuturi, 2019; Bernton et al., 2019), Sinkhorn divergence (Genevay et al., 2018), energy distance (Nguyen et al., 2020), classification accuracy (Gutmann et al., 2017), or the maximum mean discrepancy, the latter of which is the topic of this paper. Here, "efficiently estimated" is defined in terms of *sample complexity*, which is the rate of convergence at which a statistical distance can be estimated from samples. The faster an estimator converges in the number of samples, the less we need to simulate from the model, and hence, the smaller the computational cost.

We focus on the maximum mean discrepancy (MMD) (Gretton et al., 2006; 2012), a probability metric which measures the distance between distributions through the distance between their embeddings in a reproducing kernel Hilbert space; see Figure 1 for an illustration. A number of advantages of this distance are commonly put forward in the literature: (i) it has relatively low sample complexity when compared to its alternatives listed above, (ii) it has desirable statistical properties, such as leading to consistent and robust estimators, (iii) it is applicable on any data-type for which a kernel can be defined, and does not require hand-crafted summary statistics. Due to these attractive properties, the MMD has been used in a range of frameworks for likelihood-free inference, including for approximate Bayesian computation (ABC) (Park et al., 2015; Mitrovic et al., 2016; Kajihara et al., 2018; Bharti et al., 2022a; Legramanti et al., 2022), for minimum distance estimation (MDE) (Briol et al., 2019a; Chérief-Abdellatif & Alquier, 2022; Alquier & Gerber, 2020; Niu et al., 2023; Key et al., 2021), for generalised Bayesian inference (Chérief-Abdellatif & Alquier, 2020; Pacchiardi & Dutta, 2021), for Bayesian nonparametric learning (Dellaporta et al., 2022), and for training generative adversarial networks (Dziugaite et al., 2015; Li et al., 2015; 2017a; Bińkowski et al., 2018).

In this paper, we do not revisit the question of whether the MMD is the best choice of distance for a particular problem. Instead, we assume that the MMD has been chosen, and focus on constructing estimators with strong sample complexity for this distance. The most common estimators for the MMD are U-statistic or V-statistic estimators, and these have sample complexity of $\mathcal{O}(m^{-\frac{1}{2}})$, under mild conditions (Briol et al., 2019a), where m is the number of samples. In recent work, Niu et al. (2023) showed that this can be improved to $\mathcal{O}(m^{-1+\epsilon})$ for any $\epsilon > 0$ through the use of a V-statistic estimator and randomised quasi-Monte Carlo (RQMC) sampling. This significant improvement does come at the cost of restrictive assumptions — the simulator must be written in a form where the inputs are uniform random variables, and must satisfy stringent smoothness conditions which are difficult to verify in practice.

In this paper, we propose a novel set of *optimally-weighted* estimators with sample complexity of $\mathcal{O}(m^{-\frac{\nu_c}{s}-\frac{1}{2}})$ where s is the dimension of the base space and ν_c is a parameter depending on the smoothness of the kernel and the simulator. This leads to significantly improved sample complexity against both U- or V-statistic and independent samples for any ν_c, and against RQMC when $\nu_c > s/2$. Additionally, the optimality of the weights guarantees that even if this condition is not satisfied, the order of the sample complexity is still at least as good as that for existing estimators.

The remainder of the paper is structured as follows. Section 2 recalls existing estimators for the MMD, and how these are used in likelihood-free inference. Section 3 presents our estimators, and Section 4 provides a theoretical analysis of their sample complexity. Finally, Section 5 demonstrates strong empirical performance on a range of simulators, and Section 6 discusses future research.

2. Background

Throughout the paper, $\mathcal{X}$ will denote some set, and $\mathcal{P}(\mathcal{X})$ will be the set of all Borel probability measures on $\mathcal{X}$.

Likelihood-free inference. We consider the classic parameter estimation problem, where we assume that we observe some independent and identically distributed (iid) realisations $\{x_i\}_{i=1}^n \subseteq \mathcal{X}$ from some data-generating mechanism $\mathbb{Q} \in \mathcal{P}(\mathcal{X})$. Given $\{x_i\}_{i=1}^n$ and a parametric family of distributions $\{\mathbb{P}_\theta : \theta \in \Theta\} \subset \mathcal{P}(\mathcal{X})$ (i.e. the model) with parameter space Θ, we are interested in recovering the parameter value $\theta^* \in \Theta$ such that $\mathbb{P}_{\theta^*}$ is either equal, or in some sense closest, to $\mathbb{Q}$.

The challenge in likelihood-free inference is that the likelihood associated with $\mathbb{P}_\theta$ is intractable, meaning it cannot be evaluated pointwise. This prevents the use of classical methods such as maximum likelihood estimation or (exact) Bayesian inference. Instead, we assume that we are able to simulate iid realisations from $\mathbb{P}_\theta$, and such models are hence called generative models or simulator-based models. Such models are characterised through their generative process, a pair $(G_\theta, \mathbb{U})$ consisting of a simple distribution $\mathbb{U}$ (such as a multivariate Gaussian or uniform distribution) on a space $\mathcal{U}$ and a map $G_\theta : \mathcal{U} \to \mathcal{X}$ called the generator or simulator. We will call $\mathbb{U}$ a base measure and $\mathcal{U}$ the base space, and consider $\mathcal{U} \subset \mathbb{R}^s$ and $\mathcal{X} \subseteq \mathbb{R}^d$. To sample $y \sim \mathbb{P}_\theta$, one can first sample $u \sim \mathbb{U}$, then apply the generator $y = G_\theta(u)$. To perform parameter estimation for these models, it is common to repeatedly sample simulated data from the model for different parameter values and compare them to $\{x_i\}_{i=1}^n$ using a distance. We now recall the distance which will be the focus of this paper.

Maximum mean discrepancy (MMD). Let $\mathcal{H}_k$ be a reproducible kernel Hilbert space (RKHS) associated with the symmetric and positive definite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (Berlinet & Thomas-Agnan, 2004), called a reproducing

kernel, and denote by $\|\cdot\|_{\mathcal{H}_k}$ and $\langle\cdot,\cdot\rangle_{\mathcal{H}_k}$ the corresponding norm and inner product. Additionally, let $\mathcal{P}_k(\mathcal{X}) := \{\mathbb{P} \in \mathcal{P}(\mathcal{X}) : \int_{\mathcal{X}} \sqrt{k(x,x)}\mathbb{P}(\mathrm{d}x) < \infty\}$; whenever k is bounded, $\mathbb{P}_k(\mathcal{X}) = \mathbb{P}(\mathcal{X})$. As illustrated in the sketch in Figure 1, any distribution $\mathbb{P} \in \mathcal{P}_k(\mathcal{X})$ can be mapped into $\mathcal{H}_k$ via its kernel mean embedding, defined as $\mu_{k,\mathbb{P}} = \int_{\mathcal{X}} k(\cdot,x)\mathbb{P}(\mathrm{d}x)$. Then, the MMD between $\mathbb{P}$ and $\mathbb{Q}$ is the distance between their embeddings in $\mathcal{H}_k$:

$$\mathrm{MMD}_k(\mathbb{P},\mathbb{Q}) = \|\mu_{k,\mathbb{P}} - \mu_{k,\mathbb{Q}}\|_{\mathcal{H}_k}, \tag{1}$$

see Muandet et al. (2017) for a review. Alternatively, the MMD can also be expressed as $\mathrm{MMD}_k(\mathbb{P},\mathbb{Q}) = \sup_{\|f\|_{\mathcal{H}_k}\leq 1} \left|\int_{\mathcal{X}} f(x)\mathbb{P}(\mathrm{d}x) - \int_{\mathcal{X}} f(x)\mathbb{Q}(\mathrm{d}x)\right|$, where the supremum is taken over all the functions in the unit-ball of the RKHS $\mathcal{H}_k$. Whenever k is a characteristic kernel, the MMD is a probability metric, meaning that $\mathrm{MMD}_k(\mathbb{P},\mathbb{Q}) = 0$ if and only if $\mathbb{P} = \mathbb{Q}$. This condition is satisfied for kernels including the squared-exponential (SE) $k_{\mathrm{SE}}(x,y) = \eta\exp(-\|x-y\|_2^2/l^2)$, the Matérn $k_\nu(x,y) = \frac{\eta}{\Gamma(\nu)2^{\nu-1}}(\frac{\sqrt{2\nu}}{l}\|x-y\|_2)^\nu K_\nu(\frac{\sqrt{2\nu}}{l}\|x-y\|_2)$, where K_ν is the modified Bessel function of the second kind, and the inverse-multiquadric kernels on $\mathcal{X} = \mathbb{R}^d$ (Sriperumbudur et al., 2010). Matérn kernels are of particular interest: the *order* parameter ν uniquely determines the smoothness of $\mathcal{H}_k$, and for half-integer orders $\nu \in \{\frac{1}{2}, \frac{3}{2}, \dots\}$, the kernel k_ν can be written as a product of an exponential and a polynomial of order $\lfloor\nu\rfloor$ (Rasmussen & Williams, 2006).

Unfortunately, the expression in (1) usually cannot be computed directly since $\mu_{k,\mathbb{P}}$ will not be available in closed form outside of a limited number of $(k,\mathbb{P})$ pairs. Instead, using the reproducing property (i.e. $f(x) = \langle f, k(\cdot,x)\rangle_{\mathcal{H}_k}$ $\forall f \in \mathcal{H}_k$), we can write

$$\begin{aligned}\mathrm{MMD}_k^2(\mathbb{P},\mathbb{Q}) = &\int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\mathbb{P}(\mathrm{d}x)\mathbb{P}(\mathrm{d}y)\\ &- 2\int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\mathbb{P}(\mathrm{d}x)\mathbb{Q}(\mathrm{d}y)\\ &+ \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\mathbb{Q}(\mathrm{d}x)\mathbb{Q}(\mathrm{d}y).\end{aligned} \tag{2}$$

This expression is convenient to work with as it can be estimated through approximations of the integrals. Let $\{y_i\}_{i=1}^m \sim \mathbb{P}$, $\{x_i\}_{i=1}^n \sim \mathbb{Q}$ and let $\mathbb{P}^m = \frac{1}{m}\sum_{j=1}^m \delta_{y_j}$ and $\mathbb{Q}^n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, where δ_{x_i} is a Dirac measure at x_i. The squared-MMD can be approximated through a V-statistic as

$$\begin{aligned}&\mathrm{MMD}_k^2(\mathbb{P}^m,\mathbb{Q}^n) = \tfrac{1}{m^2}\textstyle\sum_{i,j=1}^m k(y_i,y_j)\\ &- \tfrac{2}{nm}\textstyle\sum_{i=1}^n\sum_{j=1}^m k(x_i,y_j) + \tfrac{1}{n^2}\sum_{i,j=1}^n k(x_i,x_j).\end{aligned}$$

This is equivalent to approximating $\mu_{k,\mathbb{P}}$ using $\mu^{\mathrm{EW}}_{k,\mathbb{P}^m}(x) = \frac{1}{m}\sum_{i=1}^m k(x,x_i)$. Alternatively, one can use an unbiased U-statistic approximation (Gretton et al., 2012). Both of these estimates can be calculated straightforwardly via evaluations of the kernel k at a computational cost $\mathcal{O}(m^2 + mn + n^2)$.

Likelihood-free inference with the MMD. The MMD has been used within a range of frameworks. In a frequentist setting, the MMD was proposed for minimum distance estimation by Briol et al. (2019a):

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} \mathrm{MMD}_k^2(\mathbb{P}_\theta,\mathbb{Q}^n). \tag{3}$$

In practice, the minimiser is computed through an optimisation algorithm, which requires evaluations of the squared-MMD or of its gradient. Such evaluations are intractable, but any estimator can be used within a stochastic optimisation algorithm. Similar optimisation problems and stochastic approximations also arise when using the MMD for generative adversarial networks (Dziugaite et al., 2015; Li et al., 2015) and for nonparametric learning (Dellaporta et al., 2022).

In a Bayesian setting, the MMD has been used to create several pseudo-posteriors by updating a prior distribution p on Θ using data. For example, the K2-ABC posterior of Park et al. (2015) is a pseudo-posterior of the form:

$$\begin{aligned}p_{\mathrm{ABC}}(\theta|x_1\dots,x_n) \propto \int\cdots\int \Pi_{j=1}^m \mathbb{1}_{\{\mathrm{MMD}_k^2(\mathbb{P}_\theta,\mathbb{Q}^n)<\varepsilon\}}(\theta)\\ p(y_j|\theta)p(\theta)\mathrm{d}y_1,\dots,\mathrm{d}y_m.\end{aligned} \tag{4}$$

where the indicator function $\mathbb{1}_{\{A\}}$ is equal to 1 if event A holds. Here, the MMD is used to determine whether a particular instance of the parametric model is within an ε distance from the data. The K2-ABC algorithm approximates this pseudo-posterior through sampling of the model $\mathbb{P}_\theta$ which leads to the use of an estimator of the squared-MMD.

Finally, the MMD has also been used for generalised Bayesian inference, where it is used to construct the MMD-Bayes posterior (Chérief-Abdellatif & Alquier, 2020)

$$p_{\mathrm{GBI}}(\theta|x_1\dots,x_n) \propto \exp(-\mathrm{MMD}_k^2(\mathbb{P}_\theta,\mathbb{Q}^n))p(\theta).$$

Once again, this pseudo-posterior is intractable, but it can be approximated through pseudo-marginal MCMC, in which case an unbiased estimator is used in place of the squared-MMD (Pacchiardi & Dutta, 2021).

Sample complexity of MMD estimators. As highlighted above, the performance of these likelihood-free inference methods relies heavily on how accurately we can estimate the MMD using samples; that is, how fast our estimator approaches $\mathrm{MMD}_k(\mathbb{P}_\theta,\mathbb{Q})$ as a function of n and m, the number of observed and simulated data points, respectively. Let $\widehat{\mathrm{MMD}}_k(\mathbb{P}_\theta^m,\mathbb{Q}^n)$ be any estimator of the MMD based on m simulated data points. Using the triangle inequality, this error can be decomposed as follows:

$$\begin{aligned}&|\mathrm{MMD}_k(\mathbb{P}_\theta,\mathbb{Q}) - \widehat{\mathrm{MMD}}_k(\mathbb{P}_\theta^m,\mathbb{Q}^n)|\\ &\leq |\mathrm{MMD}_k(\mathbb{P}_\theta,\mathbb{Q}) - \mathrm{MMD}_k(\mathbb{P}_\theta,\mathbb{Q}^n)|\\ &+ |\mathrm{MMD}_k(\mathbb{P}_\theta,\mathbb{Q}^n) - \widehat{\mathrm{MMD}}_k(\mathbb{P}_\theta^m,\mathbb{Q}^n)|\end{aligned} \tag{5}$$

where the first term describes the approximation error due to having a finite number of data points n, and the second term describes the error due to a finite number m of simulator evaluations. To understand the behaviour of the first term, we can use the following sample complexity result for the V-statistic. The proof is a direct application of the triangle inequality together with Lemma 1 in (Briol et al., 2019a).

Theorem 1. *Suppose that* $\sup_{x,x'} k(x,x') < \infty$ *and let* $\mathbb{Q}^n$ *consist of* n *iid realisations from* $\mathbb{Q} \in \mathcal{P}_k(\mathcal{X})$*. Then, for any* $\mathbb{P} \in \mathcal{P}_k(\mathcal{X})$*, we have with high probability*

$$|MMD_k(\mathbb{P},\mathbb{Q}) - MMD_k(\mathbb{P},\mathbb{Q}^n)| = \mathcal{O}(n^{-\frac{1}{2}}).$$

When $\widehat{\text{MMD}}_k(\mathbb{P}_\theta^m, \mathbb{Q}^n)$ is also a V-statistic approximation, both terms in (5) can be tackled with this result and the overall error is of size $\mathcal{O}(n^{-\frac{1}{2}} + m^{-\frac{1}{2}})$. This shows that we should take $m = \mathcal{O}(n)$ to ensure a good enough approximation of the MMD. Though this rate has the advantage of being independent of the dimension of $\mathcal{X}$, it is relatively slow in m. We therefore require a large number of simulated data points, which can be computationally expensive.

Niu et al. (2023) recently proposed an alternate approach based on randomised quasi-Monte Carlo (RQMC) (Dick et al., 2013) samples within a V-statistic. Using stronger assumptions on $\mathbb{U}, k$ and G_θ, they are able to obtain an estimator with improved sample complexity. We now state their assumptions and result below.

For $f : \mathcal{X} \to \mathbb{R}$ and a multi-index $\alpha = (\alpha_1, \dots \alpha_d) \in \mathbb{N}^d$, we denote the $|\alpha| = \sum_{i=1}^d \alpha_i$ order partial derivative $\partial^\alpha f = \partial^{|\alpha|} f / \partial^{\alpha_1} x_1 \dots \partial^{\alpha_d} x_d$ by $\partial^\alpha f$. We say $f \in C^m(\mathcal{X})$, for $m \in \mathbb{N}$, if $\partial^\alpha f$ exists and is continuous for any $|\alpha| \in [0, m]$. For two-variable $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $\partial^{\alpha,\alpha} f$ is the α-partial derivative in each variable. The norm $\|\cdot\|_{\mathcal{L}^p(\mathcal{X})}$ for $f : \mathcal{X} \to \mathbb{R}$ is defined as $\|f\|_{\mathcal{L}^p(\mathcal{X})} = (\int_{\mathcal{X}} |f(x)|^p \mathrm{d}x)^{1/p}$. The notation $a_v : b_{-v}$ represents a point $u \in [a,b]^s$ with $u_j = a_j$ for $j \in v$, and $u_j = b_j$ for $j \notin v$.

Assumption A1'. *The base space* $\mathcal{U} = [0,1]^s$*, the base measure* $\mathbb{U}$ *is uniform on* $\mathcal{U}$*, and* $\{u_i\}_{i=1}^m \subset \mathcal{U}$ *forms an RQMC point set.*

Assumption A2'. *The generator* $G_\theta : [0,1]^s \to \mathcal{X}$ *is s.t.:*

1. $\partial^{(1,\dots,1)} G_{\theta,j} \in C([0,1]^s)$ *for all* $j = 1, \dots, d$.
2. *for all* $j = 1, \dots, d$ *and* $v \in \{0,1\}^s \setminus (0,\dots,0)$*, there is a* $p_j \in [1,\infty]$*,* $\sum_{j=1}^d p_j^{-1} \leq 1$*, such that for* $g(\cdot) = \partial^v G_{\theta,j}(\cdot : 1_{-v})$ *it holds that* $\|g\|_{\mathcal{L}^{p_j}([0,1]^{|v|})} < \infty$.

Assumption A3'. *For any* $x \in \mathcal{X}$*,* $k(x,\cdot) \in C^s(\mathcal{X})$ *and* $\forall t \in \mathbb{N}^d, |t| \leq s, \sup_{x\in\mathcal{X}} \partial^{t,t} k(x,x) < C_k$ *where* C_k *is some universal constant depending only on* k.

Theorem 2. *Under A1' to A3' and* $\mathbb{Q} \in \mathcal{P}_k(\mathcal{X})$*,*

$$|MMD_k(\mathbb{P}_\theta,\mathbb{Q}) - MMD_k(\mathbb{P}_\theta^m,\mathbb{Q})| = \mathcal{O}(m^{-1+\epsilon}).$$

In this case, the second term in (5) decreases at a faster rate than the first term and the overall error decreases as $\mathcal{O}(n^{-\frac{1}{2}} + m^{-1+\epsilon})$ for any $\epsilon > 0$. As a result, (ignoring log-terms) we can take $m = \mathcal{O}(n^{-\frac{1}{2}})$, meaning a much smaller number of simulations are required. However, the technical conditions required are either very restrictive ($\mathbb{U}$ must be uniform), or will be difficult to verify in practice (the conditions on G_θ are not very interpretable and difficult to verify). Hence, the range of cases where RQMC can be applied is limited. Additionally, when both k and G_θ are smooth, faster rates can be obtained using our optimally-weighted estimator presented in the next section.

3. Optimally-Weighted Estimators

We now present our estimator, which weighs simulated data. To that end, we denote the empirical measure of the simulated data as $\mathbb{P}_\theta^{m,w} = \sum_{i=1}^m w_i \delta_{y_i}$ where $y_i = G_\theta(u_i)$, and $w_i \in \mathbb{R}$ is the weight associated with $y_i \in \mathcal{X}$ for all $i \in \{1,\dots,m\}$. Assuming for a moment that these weights are known, then we have

$$\begin{aligned}\text{MMD}_k^2(\mathbb{P}_\theta^{m,w},\mathbb{Q}^n) &= \textstyle\sum_{i,j=1}^m w_i w_j k(y_i,y_j) \\ &- \tfrac{2}{n}\textstyle\sum_{i=1}^n \sum_{j=1}^m w_j k(x_i,y_j) + \tfrac{1}{n^2}\sum_{i,j=1}^n k(x_i,x_j).\end{aligned} \tag{6}$$

Clearly, $w_i = 1/m$ for all i recovers the V-statistic approximation of the squared-MMD, but here we have additional flexibility in how to select these weights and not impose any constraints on them beyond being real-valued. To identify our choice of weights, we will make use of a tight upper bound on the approximation error.

Theorem 3. *Let* $c : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$ *be a reproducing kernel such that* $k(x,\cdot) \circ G_\theta \in \mathcal{H}_c$ *and* $\mathbb{Q} \in \mathcal{P}_k(\mathcal{X})$*. Then,* $\exists K > 0$ *independent of* $\{u_i, y_i, w_i\}_{i=1}^m$ *but dependent on* c, k *and* G_θ *such that:*

$$\begin{aligned}&|MMD_k(\mathbb{P}_\theta,\mathbb{Q}) - MMD_k(\mathbb{P}_\theta^{m,w},\mathbb{Q})| \\ &\quad \leq K \times MMD_c\left(\mathbb{U}, \textstyle\sum_{i=1}^m w_i \delta_{u_i}\right),\end{aligned}$$

Additionally, the weights minimising this upper bound can be obtained in closed-form; i.e.

$$\begin{aligned}w^* &= \underset{w\in\mathbb{R}^m}{\arg\min}\, MMD_c\left(\mathbb{U}, \textstyle\sum_{i=1}^m w_i \delta_{u_i}\right) \\ &= c(U,U)^{-1} z(U)\end{aligned} \tag{7}$$

where $z(U)_i = \mu_{c,\mathbb{U}}(u_i) = \int_{\mathcal{U}} c(u_i,u)\mathbb{U}(du)$ *is the kernel mean embedding of* $\mathbb{U}$ *in the RKHS* $\mathcal{H}_c$ *and* $(c(U,U))_{ij} = c(u_i,u_j)$ *for all* $i,j \in \{1,\dots,m\}$.

Our *optimally-weighted (OW) estimator* is the weighted estimator in (6) with the optimal weights in (7). This corresponds to estimating $\mu_{k,\mathbb{P}_\theta}$ with a weighted approximation $\mu_{k,\mathbb{P}_\theta^m}^{\text{OW}} = \sum_{i=1}^n w_i^* k(x,x_i) = \sum_{i=1}^n w_i^* k(x, G_\theta(u_i))$ where w_i^* represents the importance of $x_i = G_\theta(u_i)$ for

our approximation. To calculate these weights, we need to evaluate $\mu_{c,\mathbb{U}}$ pointwise in closed-form. The key insight is that although $\mu_{k,\mathbb{P}_\theta}$ will usually not be available in closed-form, the same is not true for $\mu_{c,\mathbb{U}}$. This is because, unlike $\mathbb{P}_\theta$, $\mathbb{U}$ is usually a simple distribution such as a uniform, Gaussian, Gamma or Poisson. Additionally, we have full flexibility in our choice of c so long as $k(x,\cdot)\circ G_\theta \in \mathcal{H}_c$. We refer to Table 1 in (Briol et al., 2019b) or the `ProbNum` Python package (Wenger et al., 2021) for a list of known closed-form kernel embeddings.

The proof of this result (see Appendix A.1) relies on two inequalities which make the overall result tight. The first is a reverse triangle inequality, which allows us to remove dependence on the true data-generating distribution $\mathbb{Q}$, a quantity which is always unknown to us. In this sense, the bound is "worst-case optimal" over $\mathbb{Q}$, a desirable property for likelihood-free inference. The second inequality allows us to use the kernel c instead of $c_\theta(u,v) = k(G_\theta(u), G_\theta(v))$ to construct our weights. Of course, this bound is attained if $c = c_\theta$ and is therefore tight. However, in practice this choice will often be infeasible due to lack of closed-form kernel embeddings $\mu_{c_\theta,\mathbb{U}}$. We therefore choose c such that the RKHS it induces contains the RKHS induced by c_θ. At a high-level, the smaller the gap between these spaces, the better the bound will be. This choice of c will be explored further through theory (in Section 4) and experiments (in Section 5).

Related methods. The optimal weights in Theorem 3 are equivalent to Bayesian quadrature (BQ) weights (Diaconis, 1988; O'Hagan, 1991; Rasmussen & Ghahramani, 2002; Briol et al., 2019b). BQ is a method for numerical integration based on Gaussian process regression (in our case with prior mean zero and prior covariance function c). We can therefore think of our estimator as performing BQ to approximate all integrals against $\mathbb{P}$ in (2). This interpretation is helpful for selecting c — the kernel should be chosen so that the corresponding Gaussian process is a good prior for the integrands in (2). This correspondence will also help us derive sample complexity results in the next section.

Our estimator minimises $\mathrm{MMD}_c\left(\mathbb{U}, \sum_{i=1}^m w_i \delta_{u_i}\right)$ over the choice of weights, but we also have flexibility over the choice of $\{u_i\}_{i=1}^m$. Unfortunately, this optimisation cannot be solved in closed-form, and is in fact usually not convex. There is a wide range of methods which have been proposed to do point selection so as to minimise an MMD with equally-weighted points. Kernel thinning (Dwivedi & Mackey, 2021), support points (Mak & Joseph, 2018) and Stein thinning (Riabiz et al., 2022) are methods based on the MMD to subsample points given a large dataset. Kernel herding (Chen et al., 2010; Bach et al., 2012) and Stein points (Chen et al., 2018; 2019) are sequential point selection methods which use an MMD as objective. In addition, similar point selection methods have also been proposed for BQ (Gunter et al., 2014; Briol et al., 2015; Belhadji et al., 2019) and these are therefore closest to our OW setting.

4. Theoretical Guarantees

Sample complexity. The following theorem establishes a sample complexity of $\mathcal{O}(m^{-\frac{\nu_c}{s}-\frac{1}{2}})$ for our optimally-weighted estimator, where ν_c is a parameter depending on the smoothness of k and G_θ. We achieve a better rate than RQMC under milder conditions, as discussed below.

Assumption A1. *The base space $\mathcal{U} \subset \mathbb{R}^s$ is bounded, open, and convex, the data space $\mathcal{X}$ is the entire $\mathbb{R}^d$ or is bounded, open, and convex. The base measure $\mathbb{U}$ has a density $f_\mathbb{U} : \mathcal{U} \to [C_\mathbb{U}, C'_\mathbb{U}]$ for some $C_\mathbb{U}$, $C'_\mathbb{U} > 0$, and $\mathbb{P}_\theta$ has a density bounded above. The point set $\{u_i\}_{i=1}^m \subset \mathcal{U}$ has a fill distance of asymptotics $h_m = \mathcal{O}(m^{-\frac{1}{s}})$, where $h_m = \sup_{u\in\mathcal{U}} \min_{i\in[1,m]} \|u - u_i\|_2$.*

Our assumptions on $\mathcal{U}$ and $\mathbb{U}$ are milder than those of A1', which requires $\mathbb{U}$ to be uniform. The assumptions on $\mathcal{X}$ and $\mathbb{P}_\theta$ are likely to hold for simulators in practice. We replace the requirement that the point set $\{u_i\}_{i=1}^m$ is RQMC with a milder assumption on the fill distance, which quantifies how far any point in $\mathcal{U}$ can get from the set $\{u_i\}_{i=1}^m$. The fill distance asymptotics is a standard assumption that ensures the coverage of $\mathcal{U}$; for example, it holds for regular grids, and in expectation for independent samples. For further examples of point sets that guarantee small fill distance, see Wynne et al. (2021).

Assumption A2. *The generator is a map $G_\theta : \mathcal{U} \to \mathcal{X}$ such that for some integer $l > s/2$, any $j \in [1,d]$ and any multi-index $\alpha \in \mathbb{N}^d$ of size $|\alpha| \leq l$, the partial derivative $\partial^\alpha G_{\theta,j}$ exists and is bounded from above.*

Assumption A2 is more interpretable and easier to check than A2' (specifically part 2) as it just requires knowing how many derivatives G_θ has. As stated in Niu et al. (2023), a simpler condition that implies A2' needs G_θ to be smooth up to the order $l \geq s$, which rules out the standard choices of $\nu \in \{\frac{1}{2}, \frac{3}{2}, \frac{5}{2}\}$ for large enough s. In contrast, we only ask that $l > s/2$.

Assumption A3. *k is a Matérn kernel on $\mathcal{X}$ of order ν_k such that $\lfloor \nu_k + d/2 \rfloor > s/2$, or an SE kernel, and c is a Matérn kernel on $\mathcal{U}$ of order $\nu_c \leq \min(\lfloor \nu_k + d/2 \rfloor, l)$.*

A3 places less restrictions on the choice of k than A3'. Although both allow for k to be the SE kernel, as a corollary of the Sobolev embedding theorem (Adams & Fournier, 2003, Theorem 4.12), A3' only holds for a Matérn k if $\lceil \nu_k \rceil \geq s+1$ (i.e. smooth k), while our lower bound on ν_k is much less restrictive. The conditions on c are needed to ensure $k(x,\cdot)\circ G_\theta \in \mathcal{H}_c$. Note that these could be weakened using the work of (Kanagawa et al., 2020; Teckentrup, 2020;

Table 1. Average and standard deviation (in parenthesis) of estimated MMD^2 ($\times 10^{-3}$) between $\mathbb{P}_\theta^m$ and $\mathbb{P}_\theta^n$ computed over 100 runs for the V-statistic and our optimally-weighted (OW) estimator. Settings: $n = 10,000$, $m = 256$.

Model	s	d	References	IID V-stat	IID OW (ours)	RQMC V-stat	RQMC OW (ours)
g-and-k	1	1	(Bharti et al., 2022b; Niu et al., 2023)	2.25 (1.52)	**0.086** (0.049)	0.060 (0.037)	**0.059** (0.037)
Two moons	2	2	(Lueckmann et al., 2021; Wiqvist et al., 2021)	2.36 (1.94)	**0.057** (0.054)	0.056 (0.044)	**0.055** (0.044)
Bivariate Beta	5	2	(Nguyen et al., 2020; Niu et al., 2023)	2.13 (1.17)	**0.555** (0.227)	0.222 (0.111)	**0.193** (0.088)
MA(2)	12	10	(Marin et al., 2011; Nguyen et al., 2020)	2.42 (0.80)	**0.705** (0.107)	0.381 (0.054)	**0.322** (0.052)
M/G/1 queue	10	5	(Pacchiardi & Dutta, 2021; Jiang, 2018)	2.52 (1.19)	**1.71** (0.568)	**0.595** (0.134)	0.646 (0.202)
Lotka-Volterra	600	2	(Briol et al., 2019a; Wiqvist et al., 2021)	2.13 (1.10)	**2.04** (0.956)	1.44 (0.955)	**1.42** (0.942)

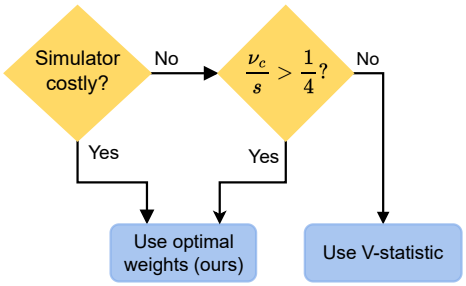


Figure 2. Guidelines on when to use our optimally-weighted estimator over the V-statistic: a) when the simulator is costly relative to the cost of MMD estimation, or , b) when ν_c is large and the dimension s is low.

Wynne et al., 2021), but at the expense of more restrictive conditions on $\{u_i\}_{i=1}^m$ in A1.

Theorem 4. *Under A1 to A3, $k(x,\cdot) \circ G_\theta \in \mathcal{H}_c$ holds, and for any and $\mathbb{Q} \in \mathcal{P}_k(\mathcal{X})$:*

$$|MMD_k(\mathbb{P}_\theta, \mathbb{Q}) - MMD_k(\mathbb{P}_\theta^{m,w}, \mathbb{Q})| = \mathcal{O}(m^{-\frac{\nu_c}{s}-\frac{1}{2}}).$$

The result shows that our method has improved sample complexity over the V-statistic for any ν_c and s. Additionally, it is better than RQMC when $\nu_c > s/2$. In practice, we should pick a kernel c that is as smooth as possible whilst not being smoother than G_θ or k, as per A3. Hence, we should take ν_c to be smaller than l and ν_k, the smoothness of G_θ and k, respectively. In case the smoothness of G_θ is unknown, the conservative choice is to take a smaller value of ν_c to ensure A3 is satisfied.

Computational Cost. The total computational cost of our method is the sum of (i) the cost of simulating from the model, which is $\mathcal{O}(mC_{\text{gen}})$, where C_{gen} is the cost of sampling one data point, and (ii) the cost of estimating MMD, which is $\mathcal{O}(m^2 + mn + n^2)$ for the V-statistic and $\mathcal{O}(m^3 + mn + n^2)$ for the OW estimator. Our method is hence slightly more expensive when m is large. However, the cost of the simulator is often the computational bottleneck, sometimes taking up to tens or hundreds of CPU hours per run; see Behrens & Dias (2015); Kirby et al. (2022). As a result, proposing data efficient likelihood-free inference methods (Beaumont et al., 2009; Gutmann & Corander, 2016; Greenberg et al., 2019) is still an active research area. In cases where $\mathcal{O}(mC_{\text{gen}}) \gg \mathcal{O}(m^3)$, the OW estimator is more efficient than the V-statistic as it requires fewer simulations to estimate the MMD. If the simulator is not more expensive than estimating the MMD and assuming a fixed computational budget, then the OW estimator achieves lower error than the V-statistic if $\nu_c/s > 1/4$ and assumptions A1 to A3 hold. This result is straightforwardly derived from Theorem 4, see Appendix A.4 for details. Figure 2 summarises the cases in which one should opt for our OW estimator instead of the V-statistic estimator.

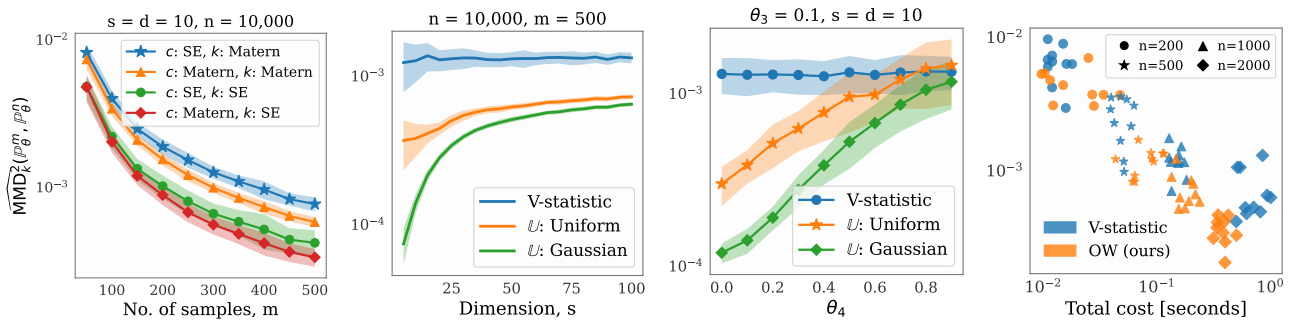
We remark that the cost of inverting the kernel matrix in our method (Equation (7)) could be reduced by using specific pairs of kernel and point sets; see Jagadeeswaran & Hickernell (2019); Karvonen et al. (2019); Karvonen & Särkkä (2019). In this case, significant gains could be observed for even cheaper simulators.

5. Numerical Experiments

We now illustrate the performance of our OW estimator on various benchmark simulators and on challenging likelihood-free inference tasks. The length-scale of kernels k and c is set using the median heuristic (Garreau et al., 2017), unless otherwise stated. The closed-form kernel mean embeddings used in the experiments are derived in Appendix A.5. Our code is available at `https://github.com/bharti-ayush/optimally-weighted_MMD`.

5.1. Benchmarking on popular simulators

We begin by comparing the V-statistic with our OW estimator on a number of popular benchmark simulators having different dimensions for $\mathcal{U} \subseteq \mathbb{R}^s$ and $\mathcal{X} \subseteq \mathbb{R}^d$. The experiments are conducted for $\{u_i\}_{i=1}^m$ being iid as well as RQMC points. We fix θ for each model (see Appendix B.1 for exact values) and estimate the MMD^2 between $\mathbb{P}_\theta^m$ and $\mathbb{P}_\theta^n$, with k and c both being the SE kernel. We set $n = 10,000$ to be large in order to make $\mathbb{P}_\theta^n$ an accurate approximation of $\mathbb{P}_\theta$, and $m = 2^8$ so as to facilitate comparison with RQMC, which requires m to be a power of 2.



(a) (b) (c) (d)

Figure 3. Error in estimating MMD^2 for the multivariate g-and-k distribution. (a) Error of our OW estimator for different choices of k and c. Increasing the smoothness of k improves the performance. (b) Comparison of V-statistic and OW estimator as a function of dimension. OW performs better for both parametrisations of $\mathbb{U}$, with the Gaussian giving lowest error. (c) Value of θ_4 also impacts the performance of the OW estimator. (d) Error vs. total computation cost for different n. OW performs better than V-statistic for similar cost: $m = n$ for V-statistic, whereas $m = (68, 126, 200, 317)$ for OW.

The results are reported in Table 1. For RQMC points, the errors are generally either similar for the two estimators (g-and-k, two moons, and Lotka-Volterra models) or smaller for the OW estimator (bivariate Beta and MA(2)), with the OW estimator achieving lower errors in all cases barring the M/G/1 queuing model. This is not surprising since the M/G/1 model has a discontinuous generator, and our theory therefore does not hold. It is also important to note that although RQMC performs very well here even without the optimal weights, the simulators were chosen in order to make this comparison feasible. In many cases, $\mathbb{U}$ will not be uniform and therefore the RQMC approach will not be possible to implement and only the iid approach is feasible.

For the iid points, the improvement in performance is much more significant. The OW estimator achieves the lowest error for all the models when $\{u_i\}_{i=1}^m$ are taken to be iid uniforms. Its error is reduced by a factor of around 20 and 40 for the g-and-k and the two moons model, respectively, compared to the V-statistic. As expected from our sample complexity results, the magnitude of this improvement reduces as s (the dimension of $\mathcal{U}$) increases. However, the OW estimator still performs slightly better than the V-statistic for the Lotka-Volterra model where $s = 600$.

5.2. Multivariate g-and-k distribution

We now assess the impact of various practical choices on the performance of our method. To do so, we consider the multivariate extension of the g-and-k distribution introduced in (Drovandi & Pettitt, 2011) and used as a benchmark in (Li et al., 2017b; Jiang, 2018; Nguyen et al., 2020). This flexible parametric family of distributions does not have a closed-form likelihood, but is easy to simulate from. We define a distribution in this family through $(G_\theta, \mathbb{U}_\theta)$, where

$$G_\theta(u) = \theta_1 + \theta_2\left[1 + 0.8\frac{1-\exp(-\theta_3 z(u))}{1+\exp(-\theta_3 z(u))}\right]\left(1+z(u)^2\right)^{\theta_4} z(u),$$

with $\theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$, $z(u) = \Sigma^{\frac{1}{2}} u$ and $\mathbb{U} = \mathcal{N}(0, I_s)$, where $\Sigma \in \mathbb{R}^{d\times d}$ is a symmetric tri-diagonal Toeplitz matrix such that $\Sigma_{ii} = 1$ and $\Sigma_{ij} = \theta_5$. The parameters θ_1,θ_2,θ_3, and θ_4 govern the location, scale, skewness, and kurtosis respectively, and $s = d$. An alternative formulation is through $(\tilde{\mathbb{U}}, \tilde{G}_\theta)$ where $\tilde{\mathbb{U}} = \text{Unif}(0,1)^s$, and $\tilde{G}_\theta = G_\theta \circ \Phi^{-1}$ where Φ is the cumulative distribution function of a $\mathcal{N}(0,1)$.

Varying choice of k and c. We first investigate the performance of our OW estimator for different combinations of k and c, the choices being either the SE or the Matérn kernel. We estimate the squared-MMD for each of these combinations as a function of m, with $d = 10$ and $n = 10,000$. The Lebesgue measure formulation is used while computing the embeddings for both the kernels. The Matérn kernel is set to order $\nu_k = \nu_c = 2.5$, and the parameter value to $\theta_0 = (3, 1, 0.1, 0.1, 0.1)$. The resulting curves are shown in Figure 3a. Our method performs best when k is the SE kernel, i.e., when it is infinitely smooth. The performance degrades slightly when k is Matérn, while the combination of c as SE and k as the Matérn kernel is the worst. This is expected from our theory, and is because the composition of G_θ and k is not smooth, but we approximate it with an infinitely smooth function. Hence, from a computational viewpoint, it is always beneficial to take k to be very smooth.

Varying dimensions s and d. We now analyse the impact of the choice of measure, either Gaussian or uniform. Figure 3b shows the OW and V-statistic estimators as the dimension $s = d$ varies. The parameter values are the same as before, $m = 500$, and the SE kernel is used for both k and

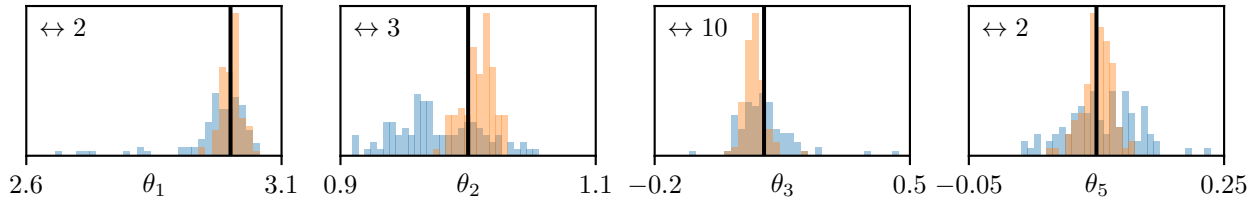


Figure 4. Histogram of parameter estimates obtained using Equation (3) with the V-statistic estimator (blue) and the OW estimator (orange) over 100 runs during the composite goodness-of-fit test. The black vertical lines denote the true value of the parameter. ↔ indicates the number of estimates for each parameter from the V-statistic that are outliers and hence not included in the plot. For the OW estimator, all estimates are within the x-axis range. The parameter estimates obtained using the OW estimator are more concentrated around the true parameter value, whereas the estimates obtained using the V-statistic have higher variance.

c. We observe that the OW estimator performs better than the V-statistic even in dimensions as high as 100. In lower dimensions, the Gaussian embedding achieves lower error than the uniform for this model, with their performance converging around $d = 60$. This is likely due to the fact that $\tilde{G}_\theta$ is an easier function to approximate than G_θ, but this is harder to assess a-priori for the user and highlights some open questions not yet covered by our theory.

Varying model parameters. Building on the previous result, we show that the performance of the OW estimation is also impacted by θ. In Figure 3c, we analyse the performance of the estimators as a function of parameter θ_4. The SE kernel is used for both k and c. While the V-statistic is not impacted by the choice of θ_4, the performance of our estimators degrade as θ_4 increases. The behaviour is similar on varying θ_3, albeit not as drastic as θ_4, see Appendix B.2 for the plot. We expect that this difference in performance is due to the regularity of the generator varying with θ.

Performance vs. computational cost. Finally, since the OW estimator tends to be more computationally expensive and this simulator is relatively cheap (≈ 1 ms to generate one sample), we also compare estimators for a fixed computational budget. To that end, we vary n and take $m = n$ for the V-statistic and $m = 2n^{2/3}$ for the OW estimator. Figure 3d shows their performance with respect to their total computational cost, including the cost of simulating from the model ($d = s = 5$). We see that the OW estimator achieves lower error on average than the V-statistic. Hence, it is preferable to use the OW estimator even for a computationally cheap simulator like the multivariate g-and-k.

Composite goodness-of-fit test. We demonstrate the performance of our method when applied to composite goodness-of-fit testing, using the method proposed by Key et al. (2021) with a test statistic based on the squared-MMD. Given iid draws from some distribution $\mathbb{Q}$, the test considers whether $\mathbb{Q}$ is an element of some parametric family $\{\mathbb{P}_\theta : \theta \in \Theta\}$ (null hypothesis) or not (alternative hypothesis). The approach uses a parametric bootstrap (Stute et al., 1993) to estimate the distribution of the squared-MMD under the null hypothesis, which can then be used to decide whether or not to reject. This requires repeatedly performing two steps: (i) estimating a parameter value through an MMD estimator of the form in Equation (3), and (ii) estimating the squared-MMD between $\mathbb{Q}$ and the model at the estimated parameter value. See Appendix B.4 for the full algorithm. This needs to be done up to B times, where B can be in the hundreds or thousands, which can be a significant challenge computationally. This limits the number of simulated samples m that can be used at each step, and is therefore a prime use case for our OW estimator.

Table 2. Fraction of repeats for which the null was rejected. An ideal test would have 0.05 when the null holds, and 1 otherwise.

Cases	IID V-stat	IID OW (ours)
$\theta_4 = 0.1$ (null holds)	0.040	0.047
$\theta_4 = 0.5$ (alternative holds)	0.040	0.413

We performed this test with a level of 0.05 using the V-statistic and OW estimator, using $B = 200$. We considered the multivariate g-and-k model with unknown $\theta_1, \theta_2, \theta_3$, and θ_5 but fixed $\theta_4 = 0.1$. We used $m = 100$ and $n = 500$ and considered two cases: $\mathbb{Q}$ is a multivariate g-and-k with $\theta_4 = 0.1$ (null holds) or $\theta_4 = 0.5$ (alternative holds). When the null hypothesis holds, we should expect the tests to reject the null at a rate close 0.05, whereas when the alternative holds, we should reject at a rate close to 1. Table 2 shows that our test based on the OW estimator performs significantly better in that respect than the V-statistic. This is due to the fact that the OW estimator is able to improve both the estimate of the parameter (see Figure 4), and the estimate of the test statistic, thus improving the overall performance.

Figure 4 shows that the estimates of the parameters computed using the OW estimator are more concentrated around the true parameter value, whereas the estimates computed using the V-statistic have higher variance. Therefore, when using the V-statistic, the distribution of the test statistic approximated by the bootstrap has higher variance, thus the estimated critical value is more conservative, and the test is

not sensitive to smaller departures from the null hypothesis. In contrast, when using the OW estimator, the estimated critical value is less conservative and the test has higher performance.

5.3. Large scale offshore wind farm model

Finally, we consider a low-order wake model (Niayifar & Porté-Agel, 2016; Kirby et al., 2023) for large-scale offshore wind farms. The model simulates an estimate of the farm-averaged local turbine thrust coefficient (Nishino, 2016), which is an indicator of the energy produced. The parameter θ is the angle (in degrees) at which the wind is blowing. The turbulence intensity is assumed to have zero-mean additive Gaussian noise (i.e. $\mathbb{U} = \mathcal{N}(0, 10^{-3})$), which then goes through the non-linear mapping of the generator. Although this model is an approximation of the state-of-the-art models that can take around 100 CPU hours per run (see e.g. (Kirby et al., 2022)), one realisation from this model takes $\approx$ 2 mins, which is still computationally prohibitive for likelihood-free inference. This example is indicative of the expensive simulators which are widely used in science, and is thus suitable for our method.

We apply the ABC method of (4) to estimate θ with both the OW estimator and the V-statistic. The tolerance threshold ε is taken in terms of percentile, i.e., the proportion of the data that yields the least MMD distances. We use 1000 parameter values from the Unif$(0, 30)$ prior on Θ. As the cost of the model far exceeds that of estimating the MMD, we take $m = 10$ for both estimators. With few m, setting the lengthscale of c using median heuristic is difficult, so we fix it to be 1. The simulated datasets took ≈ 245 hours to generate, while estimating the MMD took around 0.13 s and 0.36 s for the V-statistic and the OW estimator, respectively.

The resulting posteriors, which are approximations of the ABC posterior obtained if the MMD was computable in closed-form, are in Figure 5. We observe that the OW estimator's posterior is much more concentrated around the true value than that of the V-statistic for both values of ε. This is because the OW estimator approximates the MMD more accurately than the V-statistic for the same m. Hence, our method can achieve similar performance as the V-statistic with much smaller m, saving hours of computation time.

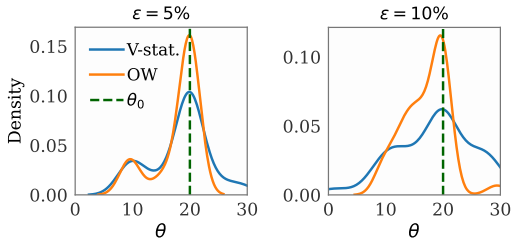


Figure 5. ABC posteriors for the wind farm model. Our OW estimator yields posterior samples that are more concentrated around the true θ_0 than the V-statistic. Performance of the U-statistic estimator is similar to the V-statistic, see Appendix B.5. Settings: $n = 100$, $\theta_0 = 20$.

6. Conclusion

We proposed an optimally-weighted MMD estimator which has improved sample complexity than the V-statistic when the generator and kernel are smooth and the dimensionality is small or moderate. Thus, our estimator requires fewer data points than alternatives in this setting, making it especially advantageous for computationally expensive simulators which are widely used in the natural sciences, biology and engineering. However, a number of open questions remain, and we highlight the most relevant below.

The parameterisation of a simulator through a generator G_θ and a measure $\mathbb{U}$ is usually not unique, and it is often unclear which parameterisation is most amenable to our method. One approach would be to choose a parameterisation where the dimension of $\mathbb{U}$ is small so as to improve the convergence rate. However, our result in Theorem 4 also contains rate constants which are difficult to get a handle on, and it is therefore difficult to identify which parameterisation is best amongst those with fixed smoothness and dimensionality.

Finally, our sample complexity result could be extended. One limitation is that we focus on the MMD and not its gradient, meaning that our results are not directly applicable for gradient-based likelihood-free inference such as the method used for our g-and-k example (Briol et al., 2019a). A future line of work could also investigate if our ideas translate to other distances used for likelihood-free inference, such as the Wasserstein distance (Bernton et al., 2019) and Sinkhorn divergence (Genevay et al., 2018; 2019).

ACKNOWLEDGEMENTS

AB was supported by the Academy of Finland (Flagship programme: Finnish Center for Artificial Intelligence FCAI). MN and OK acknowledge support from UKRI under the EPSRC grant number [EP/S021566/1]. MN was also supported through The Alan Turing Institute's Enrichment Scheme. SK was supported by the UKRI Turing AI World-Leading Researcher Fellowship, [EP/W002973/1]. FXB was supported by the Lloyd's Register Foundation Programme on Data-Centric Engineering and The Alan Turing Institute under the EPSRC grant [EP/N510129/1], and through an Amazon Research Award on "Transfer Learning for Numerical Integration in Expensive Machine Learning Systems".

References

Adams, R. A. and Fournier, J. J. *Sobolev spaces*. Elsevier, 2003.

Akeret, J., Refregier, A., Amara, A., Seehars, S., and Hasner, C. Approximate Bayesian computation for forward modeling in cosmology. *Journal of Cosmology and Astroparticle Physics*, 2015(08):043–043, 2015.

Alquier, P. and Gerber, M. Universal robust regression via maximum mean discrepancy. *arXiv:2006.00840, to appear at Biometrika*, 2020.

Aronszajn, N. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

Bach, F., Lacoste-Julien, S., and Obozinski, G. On the equivalence between herding and conditional gradient algorithms. In *Proceedings of the International Conference on Machine Learning*, pp. 1355–1362, 2012.

Beaumont, M. A. Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41(1):379–406, 2010.

Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., and Robert, C. P. Adaptive approximate Bayesian computation. *Biometrika*, 96(4):983–990, 2009.

Behrens, J. and Dias, F. New computational methods in tsunami science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 373(2053):20140382, oct 2015. doi: 10.1098/rsta.2014.0382.

Belhadji, A., Bardenet, R., and Chainais, P. Kernel quadrature with DPPs. In *Neural Information Processing Systems*, pp. 12927–12937, 2019.

Berlinet, A. and Thomas-Agnan, C. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science+Business Media, New York, 2004.

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(2):235–269, 2019. ISSN 14679868.

Bharti, A., Briol, F.-X., and Pedersen, T. A general method for calibrating stochastic radio channel models with kernels. *IEEE Transactions on Antennas and Propagation*, 70(6):3986–4001, 2022a. doi: 10.1109/tap.2021.3083761.

Bharti, A., Filstroff, L., and Kaski, S. Approximate Bayesian computation with domain expert in the loop. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 1893–1905, 2022b. URL https://proceedings.mlr.press/v162/bharti22a.html.

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. Demystifying MMD GANs. In *International Conference on Learning Representation*, 2018.

Briol, F.-X., Oates, C. J., Girolami, M., and Osborne, M. A. Frank-Wolfe Bayesian quadrature: Probabilistic integration with theoretical guarantees. In *Neural Information Processing Systems*, pp. 1162–1170, 2015.

Briol, F.-X., Barp, A., Duncan, A. B., and Girolami, M. Statistical inference for generative models with maximum mean discrepancy. *arXiv:1906.05944*, 2019a.

Briol, F.-X., Oates, C. J., Girolami, M., Osborne, M. A., and Sejdinovic, D. Probabilistic integration: A role in statistical computation? (with discussion). *Statistical Science*, 34(1):1–22, 2019b.

Cameron, E. and Pettitt, A. N. Approximate Bayesian computation for astronomical model analysis: a case study in galaxy demographics and morphological transformation at high redshift. *Monthly Notices of the Royal Astronomical Society*, 425(1):44–65, 2012. doi: 10.1111/j.1365-2966.2012.21371.x.

Chen, W. Y., Mackey, L., Gorham, J., Briol, F.-X., and Oates, C. J. Stein points. In *Proceedings of the International Conference on Machine Learning*, pp. 843–852, 2018.

Chen, W. Y., Barp, A., Briol, F.-X., Gorham, J., Girolami, M., Mackey, L., and Oates, C. J. Stein point Markov chain Monte Carlo. In *Proceedings of the International Conference on Machine Learning*, pp. 1011–1021, 2019.

Chen, Y., Welling, M., and Smola, A. Super-samples from kernel herding. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pp. 109–116, 2010.

Chérief-Abdellatif, B.-E. and Alquier, P. MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy. In *Proceesings of the 2nd Symposium on Advances in Approximate Bayesian Inference*, pp. 1–21, 2020.

Chérief-Abdellatif, B.-E. and Alquier, P. Finite sample properties of parametric MMD estimation: robustness to misspecification and dependence. *arXiv:1912.05737, to appear at Bernoulli*, 2022.

Constantine, G. M. and Savits, T. H. A multivariate faa di bruno formula with applications. *Transactions of the American Mathematical Society*, 348(2):503–520, 1996.

Cranmer, K., Brehmer, J., and Louppe, G. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

Dellaporta, C., Knoblauch, J., Damoulas, T., and Briol, F.-X. Robust Bayesian inference for simulator-based models via the MMD posterior bootstrap. In *Proceedings of the International Conference in Artificial Intelligence and Statistics*, pp. 943–970, 2022.

Diaconis, P. Bayesian Numerical Analysis. *Statistical Decision Theory and Related Topics IV*, pp. 163–175, 1988.

Dick, J., Kuo, F. Y., and Sloan, I. H. High-dimensional integration: The quasi-Monte Carlo way. *Acta Numerica*, 22(April 2013):133–288, 2013.

Drovandi, C. C. and Pettitt, A. N. Likelihood-free Bayesian estimation of multivariate quantile distributions. *Computational Statistics & Data Analysis*, 55(9):2541–2556, 2011.

Dwivedi, R. and Mackey, L. Kernel Thinning. In *Conference on Learning Theory*, volume 134, 2021.

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. Training generative neural networks via maximum mean discrepancy optimization. In *Uncertainty in Artificial Intelligence*, 2015.

Evans, L. C. and Garzepy, R. F. *Measure theory and fine properties of functions*. Routledge, 2018.

Garreau, D., Jitkrittum, W., and Kanagawa, M. Large sample analysis of the median heuristic. *arXiv:1707.07269*, 2017.

Genevay, A., Peyré, G., and Cuturi, M. Learning generative models with Sinkhorn divergences. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617, 2018.

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. Sample complexity of Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Greenberg, D., Nonnenmacher, M., and Macke, J. Automatic posterior transformation for likelihood-free inference. In *Proceedings of the International Conference on Machine Learning*, pp. 2404–2414, 2019. URL https://proceedings.mlr.press/v97/greenberg19a.html.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, volume 19, 2006.

Gretton, A., Borgwardt, K., Rasch, M. J., and Scholkopf, B. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Gunter, T., Garnett, R., Osborne, M., Hennig, P., and Roberts, S. Sampling for inference in probabilistic models with fast Bayesian quadrature. In *Advances in Neural Information Processing Systems*, pp. 2789–2797, 2014.

Gutmann, M. U. and Corander, J. Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*, 17(125): 1–47, 2016. URL http://jmlr.org/papers/v17/15-017.html.

Gutmann, M. U., Dutta, R., Kaski, S., and Corander, J. Likelihood-free inference via classification. *Statistics and Computing*, 28(2):411–425, 2017. doi: 10.1007/s11222-017-9738-6.

Hoppe, M., Embreus, O., and Fülöp, T. Dream: A fluid-kinetic framework for tokamak disruption runaway electron simulations. *Computer Physics Communications*, 268:108098, 2021. ISSN 0010-4655. doi: 10.1016/j.cpc.2021.108098. URL https://doi.org/10.1016/j.cpc.2021.108098.

Jagadeeswaran, R. and Hickernell, F. J. Fast automatic bayesian cubature using lattice sampling. *Statistics and Computing*, 29(6):1215–1229, sep 2019. doi: 10.1007/s11222-019-09895-9.

Jennings, N. R. Agent-based computing: Promise and perils. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 1999.

Jiang, B. Approximate Bayesian computation with Kullback-Leibler divergence as data discrepancy. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 1711–1721, 2018.

Kajihara, T., Yamazaki, K., Kanagawa, M., and Fukumizu, K. Kernel recursive ABC: Point estimation with intractable likelihood. In *International Conference on Machine Learning*, pp. 2400–2409, 2018.

Kanagawa, M., Sriperumbudur, B. K., and Fukumizu, K. Convergence analysis of deterministic kernel-based quadrature rules in misspecified settings. *Foundations of Computational Mathematics*, 20:155–194, 2020.

Karvonen, T. and Särkkä, S. Gaussian kernel quadrature at scaled Gauss–Hermite nodes. *BIT*, 59(4):877–902, December 2019.

Karvonen, T., Särkkä, S., and Oates, C. J. Symmetry exploits for bayesian cubature methods. *Stat. Comput.*, 29 (6):1231–1248, November 2019.

Key, O., Fernandez, T., Gretton, A., and Briol, F.-X. Composite goodness-of-fit tests with kernels. In *NeurIPS 2021 Workshop Your Model Is Wrong: Robustness and Misspecification in Probabilistic Modeling*, 2021.

Kirby, A., Nishino, T., and Dunstan, T. Two-scale interaction of wake and blockage effects in large wind farms. *arXiv:2207.03148*, 2022.

Kirby, A., Briol, F.-X., Dunstan, T. D., and Nishino, T. Data-driven modelling of turbine wake interactions and flow resistance in large wind farms. *arXiv:2301.01699*, 2023. doi: 10.48550/ARXIV.2301.01699. URL https://arxiv.org/abs/2301.01699.

Kopka, P., Wawrzynczak, A., and Borysiewicz, M. Application of the approximate Bayesian computation methods in the stochastic estimation of atmospheric contamination parameters for mobile sources. *Atmospheric Environment*, 145:201–212, 2016. doi: 10.1016/j.atmosenv.2016.09.029.

Kypraios, T., Neal, P., and Prangle, D. A tutorial introduction to Bayesian inference for stochastic epidemic models using approximate Bayesian computation. *Mathematical Biosciences*, 287:42–53, 2017. doi: 10.1016/j.mbs.2016.07.001.

Legramanti, S., Durante, D., and Alquier, P. Concentration and robustness of discrepancy-based ABC via Rademacher complexity. *arXiv:2206.06991*, 2022. URL http://arxiv.org/abs/2206.06991.

Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., and Póczos, B. MMD GAN: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, pp. 2203–2213, 2017a.

Li, J., Nott, D., Fan, Y., and Sisson, S. Extending approximate Bayesian computation methods to high dimensions via a Gaussian copula model. *Computational Statistics & Data Analysis*, 106:77–89, Feb 2017b. doi: 10.1016/j.csda.2016.07.005.

Li, Y., Swersky, K., and Zemel, R. Generative moment matching networks. In *International Conference on Machine Learning*, pp. 1718–1727, 2015.

Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S., and Corander, J. Fundamentals and recent developments in approximate Bayesian computation. *Systematic Biology*, 66:66–82, 2017.

Lueckmann, J.-M., Boelts, J., Greenberg, D., Goncalves, P., and Macke, J. Benchmarking simulation-based inference. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 343–351, 2021.

Mak, S. and Joseph, V. R. Support points. *Annals of Statistics*, 46(6A):2562–2592, 2018.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2011.

Mitrovic, J., Sejdinovic, D., and Teh, Y.-W. DR-ABC: Approximate Bayesian computation with kernel-based distribution regression. In *Proceedings of the International Conference on Machine Learning*, pp. 1482–1491, 2016. URL https://proceedings.mlr.press/v48/mitrovic16.html.

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017. doi: 10.1561/2200000060.

Nguyen, H. D., Arbel, J., Lu, H., and Forbes, F. Approximate Bayesian computation via the energy statistic. *IEEE Access*, 8:131683–131698, 2020.

Niayifar, A. and Porté-Agel, F. Analytical modeling of wind farms: A new approach for power prediction. *Energies*, 9 (9):1–13, 2016.

Niederer, S. A., Lumens, J., and Trayanova, N. A. Computational models in cardiology. *Nature Reviews Cardiology*, 16(2):100–111, 2019. ISSN 17595010. doi: 10.1038/s41569-018-0104-y. URL http://dx.doi.org/10.1038/s41569-018-0104-y.

Nishino, T. Two-scale momentum theory for very large wind farms. *Journal of Physics: Conference Series*, 753 (3), 2016.

Niu, Z., Meier, J., and Briol, F.-X. Discrepancy-based inference for intractable generative models using quasi-Monte Carlo. *Electronic Journal of Statistics*, 17(1): 1411–1456, 2023.

O'Hagan, A. Bayes-Hermite quadrature. *Journal of Statistical Planning and Inference*, 29:245–260, 1991.

Pacchiardi, L. and Dutta, R. Generalized Bayesian likelihood-free inference using scoring rules estimators. *arXiv:2104.03889*, 2021.

Park, M., Jitkrittum, W., and Sejdinovic, D. K2-ABC: approximate Bayesian computation with kernel embeddings. *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 51:398–407, 2015.

Peyré, G. and Cuturi, M. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. doi: 10.1561/2200000073.

Rasmussen, C. and Ghahramani, Z. Bayesian Monte Carlo. In *Advances in Neural Information Processing Systems*, pp. 489–496, 2002.

Rasmussen, C. and Williams, C. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Riabiz, M., Chen, W., Cockayne, J., Swietach, P., Niederer, S. A., Mackey, L., and Oates, C. J. Optimal thinning of MCMC output. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4):1059–1081, 2022.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. G. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11, 2010.

Stein, E. M. *Singular integrals and differentiability properties of functions*, volume 2. Princeton university press, 1970.

Stute, W., Manteiga, W. G., and Quindimil, M. P. Bootstrap based goodness-of-fit-tests. *Metrika*, 40(1):243–256, 1993. doi: 10.1007/BF02613687.

Teckentrup, A. L. Convergence of Gaussian process regression with estimated hyper-parameters and applications in Bayesian inverse problems. *SIAM-ASA Journal on Uncertainty Quantification*, 8(4):1310–1337, 2020.

Verbeek, M. *A Guide to Modern Econometrics, Fifth Edition*. Wiley, May 2018. ISBN 1119401151. URL https://www.ebook.de/de/product/41311245/marno_verbeek_a_guide_to_modern_econometrics_fifth_edition.html.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Wendland, H. *Scattered Data Approximation*. Cambridge University Press, 2005.

Wenger, J., Krämer, N., Pförtner, M., Schmidt, J., Bosch, N., Effenberger, N., Zenn, J., Gessner, A., Karvonen, T., Briol, F.-X., Mahsereci, M., and Hennig, P. ProbNum: Probabilistic numerics in Python. *arXiv:2112.02100*, 2021. URL http://arxiv.org/abs/2112.02100.

Wiqvist, S., Frellsen, J., and Picchini, U. Sequential neural posterior and likelihood approximation. *arXiv:2102.06522*, 2021.

Wood, S. N. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104, 2010. doi: 10.1038/nature09319.

Wynne, G., Briol, F.-X., and Girolami, M. Convergence guarantees for Gaussian process means with misspecified likelihoods and smoothness. *Journal of Machine Learning Research*, 22(123):1–40, 2021.

Supplementary Materials

In Appendix A, we present the proofs and derivations of all the theoretical results in our paper, while Appendix B contains additional details regarding our experiments.

A. Proof of Theoretical Results

In this section, we prove Theorems 3 and 4 and intermediate results required, and expand on the technical background.

A.1. Proof of Theorem 3

Proof. Let $\mathbb{P}_\theta^{m,w} = \sum_{i=1}^m w_i \delta_{y_i} = \sum_{i=1}^m w_i \delta_{G_\theta(u_i)}$. Using the fact that the MMD is a metric, we can use the reverse triangle inequality to get

$$|\text{MMD}_k(\mathbb{P}_\theta, \mathbb{Q}) - \text{MMD}_k(\mathbb{P}_\theta^{m,w}, \mathbb{Q})| \leq \text{MMD}_k(\mathbb{P}_\theta, \mathbb{P}_\theta^{m,w}).$$

Define a kernel c_θ on $\mathcal{U}$ as $c_\theta(u, u') = k(G_\theta(u), G_\theta(u'))$. As $\mathbb{P}_\theta$ is a pushforward of $\mathbb{U}$ under G_θ, it holds that:

$$\begin{aligned}
\text{MMD}_k^2(\mathbb{P}_\theta, \mathbb{P}_\theta^{m,w}) &= \int_\mathcal{X}\int_\mathcal{X} k(x,x')\mathbb{P}_\theta(\mathrm{d}x)\mathbb{P}_\theta(\mathrm{d}x') - 2\sum_{i=1}^m w_i \int_\mathcal{X} k(x_i, x)\mathbb{P}_\theta(\mathrm{d}x) + \sum_{i,j=1}^m w_i w_j k(x_i, x_j) \\
&= \int_\mathcal{X}\int_\mathcal{U} k(G_\theta(u), G_\theta(u'))\mathbb{U}(\mathrm{d}u)\mathbb{U}(\mathrm{d}u') - 2\sum_{i=1}^m w_i \int_\mathcal{U} k(G_\theta(u_i), G_\theta(u))\mathbb{U}(\mathrm{d}u) \\
&\qquad + \sum_{i,j=1}^m w_i w_j k(G_\theta(u_i), G_\theta(u_j)) \\
&= \text{MMD}_{c_\theta}^2(\mathbb{U}, \textstyle\sum_{i=1}^m w_i \delta_{u_i}).
\end{aligned}$$

Since $c_\theta(u, \cdot) \in \mathcal{H}_c$ for all $u \in \mathcal{U}$—by the assumption that $k(x, \cdot) \circ G_\theta \in \mathcal{H}_c$ for all $x \in \mathcal{X}$—it holds that $\mathcal{H}_{c_\theta} \subseteq \mathcal{H}_c$. If $\mathcal{H}_{c_\theta} = \mathcal{H}_c$, we have $\text{MMD}_k(\mathbb{P}_\theta, \mathbb{P}_\theta^{m,w}) = \text{MMD}_c(\mathbb{U}, \sum_{i=1}^m w_i \delta_{u_i})$, and the result holds for $K = 1$.

Suppose $\mathcal{H}_{c_\theta} \subset \mathcal{H}_c$. Then, by Aronszajn (1950, Theorem I.13.IV), for any $f \in \mathcal{H}_{c_\theta}$ there is a constant K independent of f such that $\|f\|_{\mathcal{H}_c} \leq K\|f\|_{\mathcal{H}_{c_\theta}}$. Together with the fact that MMD_{c_θ} is an integral-probability metric with underlying function class being the unit-ball in $\mathcal{H}_{c\theta}$, this gives

$$\begin{aligned}
\text{MMD}_{c_\theta}(\mathbb{U}, \textstyle\sum_{i=1}^m w_i \delta_{u_i}) &= \sup_{\|f\|_{\mathcal{H}_{c_\theta}} \leq 1} \left|\int_\mathcal{U} f(u)\mathbb{U}(\mathrm{d}u) - \sum_{i=1}^m w_i f(u_i)\right| \\
&= K \times \sup_{\|f\|_{\mathcal{H}_{c_\theta}} \leq 1/K} \left|\int_\mathcal{U} f(u)\mathbb{U}(\mathrm{d}u) - \sum_{i=1}^m w_i f(u_i)\right| \\
&\leq K \times \sup_{\substack{f \in \mathcal{H}_{c_\theta} \\ \|f\|_{\mathcal{H}_c} \leq 1}} \left|\int_\mathcal{U} f(u)\mathbb{U}(\mathrm{d}u) - \sum_{i=1}^m w_i f(u_i)\right| \\
&\leq K \times \sup_{\|f\|_{\mathcal{H}_c} \leq 1} \left|\int_\mathcal{U} f(u)\mathbb{U}(\mathrm{d}u) - \sum_{i=1}^m w_i f(u_i)\right| \\
&= K \times \text{MMD}_c(\mathbb{U}, \textstyle\sum_{i=1}^m w_i \delta_{u_i}),
\end{aligned}$$

where the second equality is simply a reparametrisation from f to Kf, and the inequalities use the fact that supremum of a set is not greater than supremum of its superset, and

$$\{f \in \mathcal{H}_{c_\theta} \mid K\|f\|_{\mathcal{H}_{c_\theta}} \leq 1\} \subseteq \{f \in \mathcal{H}_{c_\theta} \mid \|f\|_{\mathcal{H}_c} \leq 1\} \subseteq \{f \in \mathcal{H}_c \mid \|f\|_{\mathcal{H}_c} \leq 1\}.$$

Note that the tightness of the bound will depend on the gap between $\mathcal{H}_{c_\theta}$ and $\mathcal{H}_c$; the smaller this gap, the tighter the bound will be. This is illustrated in Figure 6.

To prove the result about the exact form of w, we note that

$$\underset{w \in \mathbb{R}^m}{\arg\min}\, \text{MMD}_c\left(\mathbb{U}, \textstyle\sum_{i=1}^m w_i \delta_{u_i}\right) = \underset{w \in \mathbb{R}^m}{\arg\min}\, \text{MMD}_c^2\left(\mathbb{U}, \textstyle\sum_{i=1}^m w_i \delta_{u_i}\right),$$

and

$$\text{MMD}_c^2\left(\mathbb{U}, \textstyle\sum_{i=1}^m w_i \delta_{u_i}\right) = \int_\mathcal{U}\int_\mathcal{U} c(u,v)\mathbb{U}(\mathrm{d}u)\mathbb{U}(\mathrm{d}v) - 2\sum_{i=1}^m w_i \int_\mathcal{U} c(u_i, u)\mathbb{U}(\mathrm{d}u) + \sum_{i,j=1}^m w_i w_j c(u_i, u_j).$$

The latter is a quadratic form in w, meaning it can be minimised in closed-form over w and the optimal weights are given by w^*. This completes the proof of the second part of the theorem.

□

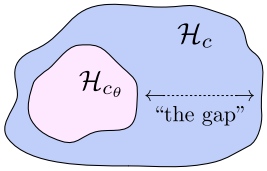


Figure 6. Pictorial representation of the gap between the RKHS induced by the kernels c and $c_\theta = k(G_\theta(u), G_\theta(v))$. The size of the gap affects the tightness of the bound in Theorem 3, and consequently Theorem 4.

A.2. Chain rule in Sobolev spaces

The proof of Theorem 4, specifically the result $k(x, \cdot) \circ G_\theta \in \mathcal{H}_c$ for Matérn k and c, will use a specific form of a chain rule for Sobolev spaces. We justify the choice of Matérn kernels—or more generally, kernels the RKHS of which is norm-equivalent to the well-studied Sobolev space—and prove the form of the chain rule for Sobolev spaces that will imply $k(x, \cdot) \circ G_\theta \in \mathcal{H}_c$.

For general c and k, $k(x, \cdot) \circ G_\theta \in \mathcal{H}_c$ is non-trivial to check. Here, we introduce sufficient conditions on c, k, and G_θ that are easily interpretable and correspond to common practical settings. Specifically, we consider c and k the RKHS of which, $\mathcal{H}_c$ and $\mathcal{H}_k$, are Sobolev spaces, and G_θ of a certain degree of smoothness—which reduces the problem to a form of a chain rule for Sobolev spaces.[1] The rest of the section proceeds as follows: first, we introduce the background definitions and results, then show that the required form of the chain rule holds for first order derivatives (Lemma 1), and finally extend the result to higher order derivatives (Theorem 6).

Background. We consider the well-studied Sobolev kernels (see e.g. Wendland, 2005, Chapter 10), which are kernels that induce a reproducing kernel Hilbert space (RKHS) that is norm-equivalent to a Sobolev space $\mathcal{W}^{l,2}(\mathcal{X})$, $\mathcal{X} \subseteq \mathbb{R}^d$, for some integer $l > d/2$. We give the definition of $\mathcal{W}^{l,2}$ Sobolev spaces below, and refer to Adams & Fournier (2003) for an in-depth treatment of Sobolev spaces and Berlinet & Thomas-Agnan (2004) for general RKHS theory.

Definition 1 (Sobolev spaces)**.** *Suppose $\mathcal{X}$ is an open subset of $\mathbb{R}^d$. The Sobolev space $\mathcal{W}^{l,2}(\mathcal{X})$, $l > d/2$, is a space of functions $f : \mathcal{X} \to \mathbb{R}$ such that $\|f\|^2_{\mathcal{L}^2(\mathcal{X})} = \int_{\mathcal{X}} f^2(x)dx < \infty$, and for any multi-index $\alpha \in \mathbb{N}^d$ with $|\alpha| = \sum_{i=1}^d \alpha_i \leq l$, the weak derivative $D^\alpha f = D^{\alpha_1}_{x_1} \dots D^{\alpha_d}_{x_d} f$ exists and $\|D^\alpha f\|_{\mathcal{L}^2(\mathcal{X})} < \infty$.*

A weak derivative is a generalisation of the concept of a derivative to functions that are not differentiable. A locally integrable function $D_{x_i} f$ is a weak derivative of f in x_i if it closely resembles the behavior of the ordinary derivative on any open $U \subseteq \mathcal{X}$: for any infinitely continuously differentiable function with a compact support, the integration chain rule holds with f and $D_{x_i} f$—as it would for an ordinary derivative. As the definition is only concerned with equality of the integrals in the chain rule, a weak derivative is not uniquely defined: two functions g_1 and g_2 can be weak derivatives of f in x_i if (and only if) they only differ on a zero-volume set, meaning a set the Lebesgue measure of which is zero. As such, by $D_{x_i} f$ we will refer to any function that satisfies the definition of a weak derivative. For a multi-index $\alpha = (\alpha_1, \dots \alpha_d) \in \mathbb{N}^d$, by $D^\alpha f$ we denote the $|\alpha|$ order weak derivative $D^\alpha f = D^{\alpha_1}_{x_1} \dots D^{\alpha_d}_{x_d} f$, where

$$D^n x_i f = \underbrace{D_{x_i} \dots D_{x_i}}_{n} f \text{ for any } n \in \mathbb{N}.$$

If an ordinary derivative $\partial^\alpha f = \partial^{|\alpha|} f / \partial^{\alpha_1} x_1 \dots \partial^{\alpha_d} x_d$ exists, it is equal to any weak $D^\alpha f$. It is important to clarify that the definition of Sobolev spaces given here is specific to the case $\mathcal{W}^{l,2}(\mathcal{X})$, $l > d/2$. General Sobolev spaces $W^{l,p}(\mathcal{X})$ are subspaces of more general Lebesgue spaces, and are spaces not of functions, but of *equivalence classes* of functions. Two functions f_1, f_2 are in the same equivalence class $[f]$ if they are equal almost everywhere. General Lebesgue and Sobolev space theory requires careful handling of the notion of equivalence classes, as the functions in them may differ arbitrarily on sets of Lebesgue measure zero. However, by Sobolev embedding theorem (Adams & Fournier, 2003, Theorem 4.12) every element of $\mathcal{W}^{l,2}(\mathcal{X})$ is continuous if $l > d/2$, which implies that every equivalence class contains exactly one function—and we may define $\mathcal{W}^{l,2}(\mathcal{X})$ as a space of functions, as is done above.

Throughout the proofs, we will say $f \in \mathcal{L}^\infty(\mathcal{X})$ if it is bounded on $\mathcal{X}$, and $f \in C^m(\mathcal{X})$, for $m \in \mathbb{N}$, if $\partial^\alpha f$ exists and is

[1] Though various forms of the chain rule for Sobolev spaces exist in the literature (for example, Evans & Garzepy (2018, Section 4.2.2)), they tend to either consider $F \circ f$, where f is in the Sobolev space (rather than F), or place overly strong assumptions on f.

continuous for any $|\alpha| \in [0, m]$. Specifically, $C^0(\mathcal{X})$ is the space of continuous functions, and $C^\infty(\mathcal{X})$ a space of infinitely differentiable functions with continuous derivatives. The output space of functions in both $\mathcal{L}^\infty(\mathcal{X})$ and $C^m(\mathcal{X})$ is omitted from the notation as it will be clear from the specific f in question.

We start by recalling an important result that characterises Sobolev functions as limit points of sequences of $C^\infty(\mathcal{X})$ functions. Since it is a necessary and sufficient condition, we will use this result both to operate on a function in a Sobolev space using the "friendlier" smooth functions, and to prove a function of interest lies in a Sobolev space by finding a sequence of smooth function that approximates it accordingly.

Theorem 5 (Theorem 3.17, (Adams & Fournier, 2003))**.** *For an open set $\mathcal{X} \subseteq \mathbb{R}^d$, a function $f : \mathcal{X} \to \mathbb{R}$ lies in the Sobolev space $\mathcal{W}^{1,2}(\mathcal{X})$ and has weak derivatives $D_{x_j}[f]$, $j \in [1, d]$ if and only if there exists a sequence of functions $f_n \in C^\infty(\mathcal{X}) \cap \mathcal{W}^{1,2}(\mathcal{X})$ such that for $j \in [1, d]$*

$$\|f - f_n\|_{\mathcal{L}^2(\mathcal{X})} \to 0, \quad n \to \infty, \tag{8}$$

$$\left\| D_{x_j}[f] - \tfrac{\partial f_n}{\partial x_j} \right\|_{\mathcal{L}^2(\mathcal{X})} \to 0, \quad n \to \infty, \tag{9}$$

where $\frac{\partial f_n}{\partial x_j}$ is the ordinary derivative of f_n with respect to x_j.

Note that the functions f_n converge to f in the Sobolev $\mathcal{W}^{1,2}(\mathcal{X})$ norm, $\|f - f_n\|^2_{\mathcal{W}^{1,2}(\mathcal{X})} = \|f - f_n\|_{\mathcal{L}^2(\mathcal{X})} + \sum_{j=1}^d \|D_{x_j} f - \partial f_n / \partial x_j\|_{\mathcal{L}^2(\mathcal{X})} \to 0$ as $n \to \infty$, if and only if (8) and (9) hold.

Chain rule for $\mathcal{W}^{1,2}$. We now prove that chain rule holds for $\varphi \circ G_\theta$ for φ in a Sobolev space $W^{1,2}(\mathcal{X})$. For clarity, we will explicitly state the assumptions on G_θ in the main text. Recall that a measure $\mathbb{P}_\theta$ on $\mathcal{X} \subseteq \mathbb{R}^d$ is said to be a pushforward of a measure $\mathbb{U}$ on $\mathcal{U} \subseteq \mathbb{R}^s$ under $G_\theta : \mathcal{U} \to \mathcal{X}$ if for any $\mathcal{X}$-measurable $f : \mathcal{X} \to \mathbb{R}$ it holds that $\int_{\mathcal{X}} f(x) \mathbb{P}_\theta(\mathrm{d}x) = \int_{\mathcal{U}} [f \circ G_\theta](u) \mathbb{U}(\mathrm{d}u)$.

Lemma 1 (Chain rule for $\mathcal{W}^{1,2}$)**.** *Suppose*

- $\varphi \in \mathcal{W}^{1,2}(\mathcal{X})$.
- *$\mathcal{U} \subset \mathbb{R}^s$ is bounded, $\mathcal{X} \subset \mathbb{R}^d$ is open, and $\mathcal{X} = G_\theta(\mathcal{U})$ for some $G_\theta = (G_{\theta,1}, \ldots, G_{\theta,d})^\top$. The partial derivative $\partial G_{\theta,j} / \partial u_i$ exists and $|\partial G_{\theta,j} / \partial u_i| \leq C_G$ for some C_G for all $i \in [1, s]$ and $j \in [1, d]$.*
- *$\mathbb{U}$ is a probability distribution on $\mathcal{U}$ that has a density $f_\mathbb{U} : \mathcal{U} \to [C_\mathbb{U}, \infty)$ for $C_\mathbb{U} > 0$.*
- *$\mathbb{P}_\theta$ is a pushforward of $\mathbb{U}$ under G_θ, and has a density $f_{\mathbb{P}_\theta}$ such that $f_{\mathbb{P}_\theta}(x) \leq C_{\mathbb{P}_\theta}$ for all $x \in \mathcal{X}$ for some $C_{\mathbb{P}_\theta}$.*

Then $\varphi \circ G_\theta \in \mathcal{W}^{1,2}(\mathcal{U})$, and for $i \in [1, s]$, its weak derivative $D_{u_i}[\varphi \circ G_\theta]$ is equal to $\sum_{j=1}^d [D_{x_j} \varphi \circ G_\theta] \frac{\partial G_{\theta,j}}{\partial u_i}$.

Proof. Since $\mathcal{X}$ is open, by Theorem 5 there is a sequence $\varphi_n \in C^\infty(\mathcal{X}) \cap \mathcal{W}^{1,2}(\mathcal{X})$ such that

$$\|\varphi - \varphi_n\|_{\mathcal{L}^2(\mathcal{X})} \to 0, \quad n \to \infty,$$
$$\left\| D_{x_j} \varphi - \tfrac{\partial \varphi_n}{\partial x_j} \right\|_{\mathcal{L}^2(\mathcal{X})} \to 0, \quad n \to \infty,$$

The proof proceeds as follows: we show that the sequence $\varphi_n \circ G_\theta$ approximates $\varphi \circ G_\theta$, and $\frac{\partial [\varphi_n \circ G_\theta]}{\partial u_i}$ approximates the sum in the statement of the lemma, $\sum_{j=1}^d [D_{x_j} \varphi \circ G_\theta] \frac{\partial G_{\theta,j}}{\partial u_i}$, in $\mathcal{L}^2(\mathcal{U})$–norm. Then, by the sufficient condition in Theorem 5, $\varphi \circ G_\theta$ lies in $\mathcal{W}^{1,2}(\mathcal{U})$, and its weak derivative in u_i is $\sum_{j=1}^d [D_{x_j} \varphi \circ G_\theta](u) \frac{\partial G_{\theta,j}}{\partial u_i}(u)$, for any $i \in [1, s]$.

Since $\mathbb{P}_\theta$ has a density, for any $\mathcal{X}$-measurable f it holds that

$$\int_{\mathcal{X}} f(x) f_{\mathbb{P}_\theta}(x) \mathrm{d}x = \int_{\mathcal{U}} [f \circ G_\theta](u) f_\mathbb{U}(u) \mathrm{d}u.$$

Together with density bounds, this gives $\|\varphi \circ G_\theta - \varphi_n \circ G_\theta\|_{\mathcal{L}^2(\mathcal{U})} \to 0$ as

$$\begin{aligned} \int_{\mathcal{U}} (\varphi \circ G_\theta(u) - \varphi_n \circ G_\theta(u))^2 \mathrm{d}u \leq C_\mathbb{U}^{-1} \int_{\mathcal{U}} (\varphi \circ G_\theta(u) - \varphi_n \circ G_\theta(u))^2 f_\mathbb{U}(u) \mathrm{d}u &= C_\mathbb{U}^{-1} \int_{\mathcal{X}} (\varphi(x) - \varphi_n(x))^2 f_{\mathbb{P}_\theta}(x) \mathrm{d}x \\ &\leq C_\mathbb{U}^{-1} C_{\mathbb{P}_\theta} \int_{\mathcal{X}} (\varphi(x) - \varphi_n(x))^2 \mathrm{d}x. \end{aligned}$$

In the same fashion, $\|D_{x_j}\varphi \circ G_\theta - \frac{\partial \varphi_n}{\partial x_j} \circ G_\theta\|_{\mathcal{L}^2(\mathcal{U})} \to 0$ since

$$\begin{aligned}\int_{\mathcal{U}} \left(D_{x_j}\varphi \circ G_\theta(u) - \tfrac{\partial \varphi_n}{\partial x_j} \circ G_\theta(u)\right)^2 \mathrm{d}u &\leq C_{\mathbb{U}}^{-1} \int_{\mathcal{U}} \left(D_{x_j}\varphi \circ G_\theta(u) - \tfrac{\partial \varphi_n}{\partial x_j} \circ G_\theta(u)\right)^2 f_{\mathbb{U}}(u)\mathrm{d}u \\ &= C_{\mathbb{U}}^{-1} \int_{\mathcal{X}} \left(D_{x_j}\varphi(x) - \tfrac{\partial \varphi_n}{\partial x_j}(x)\right)^2 f_{\mathbb{P}_\theta}(x)\mathrm{d}x \\ &\leq C_{\mathbb{U}}^{-1} C_{\mathbb{P}_\theta} \int_{\mathcal{X}} \left(D_{x_j}\varphi(x) - \tfrac{\partial \varphi_n}{\partial x_j}(x)\right)^2 \mathrm{d}x.\end{aligned}$$

Since φ and G_θ are both differentiable, the ordinary chain rules applies to $\varphi_n \circ G_\theta$,

$$\tfrac{\partial[\varphi_n \circ G_\theta]}{\partial u_i} = \textstyle\sum_{j=1}^d \left[\tfrac{\partial \varphi_n}{\partial x_j} \circ G_\theta\right] \tfrac{\partial G_{\theta,j}}{\partial u_i},$$

and for any $i \in [1, s]$ the convergence of derivatives $\|[D_{x_j}\varphi \circ G_\theta]\frac{\partial G_{\theta,j}}{\partial u_i} - \frac{\partial[\varphi_n \circ G_\theta]}{\partial u_i}\|_{\mathcal{L}^2(\mathcal{U})} \to 0$ follows since

$$\begin{aligned}\int_{\mathcal{U}} \left(\textstyle\sum_{j=1}^d \left[D_{x_j}\varphi \circ G_\theta\right] \tfrac{\partial G_{\theta,j}}{\partial u_i} - \tfrac{\partial[\varphi_n \circ G_\theta]}{\partial u_i}\right)^2 \mathrm{d}u &= \int_{\mathcal{U}} \left(\textstyle\sum_{j=1}^d \left[D_{x_j}\varphi \circ G_\theta - \tfrac{\partial \varphi_n}{\partial x_j} \circ G_\theta\right] \tfrac{\partial G_{\theta,j}}{\partial u_i}\right)^2 \mathrm{d}u \\ &\leq 2\textstyle\sum_{j=1}^d \int_{\mathcal{U}} \left(\left[D_{x_j}\varphi \circ G_\theta - \tfrac{\partial \varphi_n}{\partial x_j} \circ G_\theta\right] \tfrac{\partial G_{\theta,j}}{\partial u_i}\right)^2 \mathrm{d}u \\ &\leq 2C_G^2 \textstyle\sum_{j=1}^d \int_{\mathcal{U}} \left(D_{x_j}\varphi \circ G_\theta - \tfrac{\partial \varphi_n}{\partial x_j} \circ G_\theta\right)^2 \mathrm{d}u\end{aligned}$$

where the first inequality is using the inequality $(\sum_{i=1}^d a_i)^2 \leq 2\sum_{i=1}^d a_i^2$. This completes the proof.

□

Chain rule for $\mathcal{W}^{l,2}$. To extend Lemma 1 to Sobolev spaces of order higher than 1, we need the following version of the weak derivative product rule, for a product of a function f in $\mathcal{W}^{1,2}$ and bounded differentiable function g with bounded derivatives. Other versions of the product rule—for different regularity assumptions on g—exist in the literature (for example, Adams & Fournier (2003)); we will require this specific form.

Lemma 2 (Product rule)**.** *Suppose $\mathcal{X} \subseteq \mathbb{R}^d$ is open, $f \in \mathcal{W}^{1,2}(\mathcal{X})$, $g(x)$ is differentiable on $\mathcal{X}$, and $g(x) \leq L$, $[\partial g/\partial x_i](x) \leq L$ for all $x \in \mathcal{X}$ for some constant L. Then $fg \in \mathcal{W}^{1,2}(\mathcal{X})$ and for any $i \in [1, d]$,*

$$D_{x_i}[fg] = [D_{x_i}f]g + f\left[\partial g/\partial x_i\right]$$

Proof. By the criterion in Theorem 5, there is a sequence of smooth functions f_n approximating f, meaning

$$\begin{aligned}\int_{\mathcal{X}} (f(x) - f_n(x))^2 \mathrm{d}x &\to 0 \text{ as } n \to \infty, \\ \int_{\mathcal{X}} \left(D_{x_i}f(x) - [\partial f_n/\partial x_i](x)\right)^2 \mathrm{d}x &\to 0 \text{ as } n \to \infty.\end{aligned}$$

We will show that $f_n g$ approximates fg with weak derivatives taking the form $[D_{x_i}f]g + f[\partial g/\partial x_i]$; by the aforementioned criterion, it will follow that $fg \in \mathcal{W}^{1,2}(\mathcal{X})$.

First, we establish convergence of functions. As $n \to \infty$,

$$\|fg - f_n g\|_{\mathcal{L}^2(\mathcal{X})}^2 = \int_{\mathcal{X}} \left(f(x)g(x) - f_n(x)g(x)\right)^2 \mathrm{d}x \leq L^2 \int_{\mathcal{X}} \left(f(x) - f_n(x)\right)^2 \mathrm{d}x \to 0.$$

By the ordinary chain rule, $\partial[f_n g]/\partial x_i = [\partial f_n/\partial x_i]g + f[\partial g/\partial x_i]$. Then, applying triangle inequality for norms and the fact that $(a+b)^2 \leq 2a^2 + 2b^2$ for any a, b, we get that for $n \to \infty$,

$$\begin{aligned}\left\|\tfrac{\partial f_n}{\partial x_i} g + f_n \tfrac{\partial g}{\partial x_i} - [D_{x_i}f]\, g - f\tfrac{\partial g}{\partial x_i}\right\|_{\mathcal{L}^2(\mathcal{X})}^2 &\leq 2\left\|\tfrac{\partial f_n}{\partial x_i} g - [D_{x_i}f]\, g\right\|_{\mathcal{L}^2(\mathcal{X})}^2 + 2\left\|f_n \tfrac{\partial g}{\partial x_i} - f\tfrac{\partial g}{\partial x_i}\right\|_{\mathcal{L}^2(\mathcal{X})}^2 \\ &\leq 2L^2 \left\|\tfrac{\partial f_n}{\partial x_i} - [D_{x_i}f]\right\|_{\mathcal{L}^2(\mathcal{X})}^2 + 2L^2 \|f_n - f\|_{\mathcal{L}^2(\mathcal{X})} \to 0.\end{aligned}$$

This completes the proof. □

We are now ready to extend the chain rule from order 1—proven in Lemma 1—to arbitrary order l.

Theorem 6 (Chain rule for $\mathcal{W}^{l,2}$)**.** *Suppose*

- $\varphi \in \mathcal{W}^{l_\varphi,2}(\mathcal{X})$.
- $\mathcal{U} \subset \mathbb{R}^s$ *is bounded,* $\mathcal{X} \subset \mathbb{R}^d$ *is open, and* $\mathcal{X} = G_\theta(\mathcal{U})$ *for some* $G_\theta = (G_{\theta,1}, \dots, G_{\theta,d})^\top$*. For some* l_G *and any* $|\alpha| \leq l_G$*,* $j \in [1, s]$*, the derivative* $\partial^\alpha G_{\theta,j}$ *exists and is in* $\mathcal{L}^\infty(\mathcal{U})$*.*
- $\mathbb{U}$ *is a probability distribution on* $\mathcal{U}$ *that has a density* $f_\mathbb{U} : \mathcal{U} \to [C_\mathbb{U}, \infty)$ *for* $C_\mathbb{U} > 0$*.*
- $\mathbb{P}_\theta$ *is a pushforward of* $\mathbb{U}$ *under* G_θ *with a density bounded above.*

Then $\varphi \circ G_\theta \in \mathcal{W}^{l,2}(\mathcal{U})$ *for* $l = \min\{l_\varphi, l_G\}$*, and for any* $k \leq l$ *and* $|\alpha_0| = k$*, the derivative takes an* α_0*-specific* $(\kappa, \beta, \alpha, \eta)$*–form*

$$D^{\alpha_0}[\varphi \circ G_\theta] = \sum_{i=1}^{I} \sum_{j=1}^{d^{\kappa_i}} \left[D^{\beta_{ij}} \varphi \circ G_\theta\right] \prod_{l=1}^{\kappa_i} \partial^{\alpha_{ijl}} G_{\theta,\eta_{ijl}}, \tag{10}$$

where $I \in \mathbb{N}$*, and for any* $i \in [1, I]$*,* $k \geq \kappa_i \in \mathbb{N}$*;* $\beta_{ij} \in \mathbb{N}^d$ *is a multi-index of size* κ_i *for* $j \in [1, d^{\kappa_i}]$*;* $\alpha_{ijl} \in \mathbb{N}^s$ *is of size* $|\alpha_{ijl}| \leq k$*, and* $\eta_{ijl} \in [1, d]$ *for* $l \in [1, \kappa_i]$*.*

By saying the $(\kappa, \beta, \alpha, \eta)$ form is α_0-specific, we mean that the values of $I, (\kappa, \beta, \alpha, \eta)$ depend on α_0, and may be different for $\alpha_0' \neq \alpha_0$; we do not index $I, (\kappa, \beta, \alpha, \eta)$ by α_0 for the sake of readability.

Before proving this result, let us point out that the $(\kappa, \beta, \alpha, \eta)$–form introduced in the theorem can be seen as a form of the Faà di Bruno's formula which generalises the chain rule to higher derivatives (Constantine & Savits, 1996, Theorem 1). However, since our ultimate goal is to show $\varphi \circ G_\theta \in \mathcal{W}^{l,2}(\mathcal{U})$, and the expression for the derivative is simply a means for proving that, an unspecified $(\kappa, \beta, \alpha, \eta)$–form suffices. It is simpler to conduct a proof for general $(\kappa, \beta, \alpha, \eta)$ without using explicit Faà di Bruno forms.

Proof of Theorem 6. Note that $\varphi \circ G_\theta \in \mathcal{W}^{l,2}(\mathcal{U})$ if and only if $\varphi \circ G_\theta \in \mathcal{W}^{k,2}(\mathcal{U})$ for $k \leq l$. We use this to construct a proof by induction: we show the statement holds for $k = 1$, and that $\varphi \circ G_\theta \in \mathcal{W}^{k,2}(\mathcal{U})$ implies $\varphi \circ G_\theta \in \mathcal{W}^{k+1,2}(\mathcal{U})$ if $k + 1 \leq l$ (and the weak derivatives take a $(\kappa, \beta, \alpha, \eta)$-form stated in Equation (10)).

Case $k = 1$**:** $\varphi \circ G_\theta$ is in $\mathcal{W}^{1,2}(\mathcal{U})$.

Suppose $\alpha_0 = e[m]$ for some unit vector $e[m] = (0, \dots, 0, 1, 0, \dots, 0)$ where the 1 is the m'th element. Then, as proven in Lemma 1, $D^{e[m]}[\varphi \circ G_\theta] = D_{u_m}[\varphi \circ G_\theta]$ is equal to $\sum_{j=1}^{d} [D_{x_j} \varphi \circ G_\theta][\partial G_{\theta,j} / \partial u_m] = \sum_{j=1}^{d} [D^{e[j]} \varphi \circ G_\theta] \partial^{e[m]} G_{\theta,j}$, so the statement holds for $I = 1$, $\kappa_1 = 1$, $\beta_{1j} = e[j]$, $\alpha_{1j1} = e[m]$, $\eta_{1j1} = j$.

Case k **implies** $k + 1$**:** If $k + 1 \leq l$ and $\varphi \circ G_\theta$ is in $\mathcal{W}^{k,2}(\mathcal{U})$, and for every $|\alpha_0| = k$ Equation (10) holds for some α_0-specific $(\kappa, \beta, \alpha, \eta)$, then $\varphi \circ G_\theta$ is in $\mathcal{W}^{k+1,2}(\mathcal{U})$, and for any $|\tilde{\alpha}_0| = k + 1$ there is a $(\tilde{\kappa}, \tilde{\beta}, \tilde{\alpha}, \tilde{\eta})$–form, $|\tilde{\kappa}| = \tilde{I}$,

$$D^{\tilde{\alpha}_0}[\varphi \circ G_\theta] = \sum_{i=1}^{\tilde{I}} \sum_{j=1}^{d^{\tilde{\kappa}_i}} \left[D^{\tilde{\beta}_{ij}} \varphi \circ G_\theta\right] \prod_{l=1}^{\tilde{\kappa}_i} \partial^{\tilde{\alpha}_{ijl}} G_{\theta,\tilde{\eta}_{ijl}}. \tag{11}$$

By induction assumption, $\varphi \circ G_\theta$ is in $\mathcal{W}^{k,2}(\mathcal{U})$, so it is in $\mathcal{W}^{k+1,2}(\mathcal{U})$ if and only if $D^{\alpha_0}[\varphi \circ G_\theta]$ is in $\mathcal{W}^{1,2}(\mathcal{U})$ for any α_0 of size k. The latter can be shown by studying the $(\kappa, \beta, \alpha, \eta)$–form that $D^{\alpha_0}[\varphi \circ G_\theta]$ takes by (10), for some α_0–specific $(\kappa, \beta, \alpha, \eta)$. Since $l_\varphi \geq l \geq k + 1$ (the last inequality holds by the induction assumption), it holds that $\mathcal{W}^{l_\varphi,2}(\mathcal{X}) \subseteq \mathcal{W}^{l,2}(\mathcal{X}) \subseteq \mathcal{W}^{k+1,2}(\mathcal{X})$. Then $\varphi \in \mathcal{W}^{k+1,2}(\mathcal{X})$, and since $|\beta_{ij}| = \kappa_i \leq k$ by definition of β_{ij}, we have $D^{\beta_{ij}} \varphi \in \mathcal{W}^{1,2}(\mathcal{X})$ for all i, j. Then by Lemma 1, its composition with G_θ is in $\mathcal{W}^{1,2}(\mathcal{U})$, meaning $D^{\beta_{ij}} \varphi \circ G_\theta \in \mathcal{W}^{1,2}(\mathcal{U})$. Consequently, $D^{\alpha_0}[\varphi \circ G_\theta]$ as per Equation (10) is a sum over the product of functions in $\mathcal{W}^{1,2}(\mathcal{U})$, and bounded functions with bounded derivatives; by Lemma 2, such product is in $\mathcal{W}^{1,2}(\mathcal{U})$, and it follows that $D^{\alpha_0}[\varphi \circ G_\theta] \in \mathcal{W}^{1,2}(\mathcal{U})$ as well.

Finally, we show that for any fixed $|\tilde{\alpha}_0| = k + 1$ there are $\tilde{I}, \tilde{\kappa}, \tilde{\beta}, \tilde{\alpha}, \tilde{\eta}$ for which (11) holds; this will conclude the induction step. Suppose α_0 of size k, $|\alpha_0| = k$, is such that $\tilde{\alpha}_0 = \alpha_0 + e[m]$ for some α_0 (that is unrelated to α_0 in the previous part of the proof) and a unit vector $e[m]$ (such pair of m and α_0 must exist as $|\tilde{\alpha}_0| = k + 1$). For this α_0, in a slight abuse of

notation, we shall say that $\kappa, \beta, \alpha, \eta$ are such that $D^{\alpha_0}[\varphi \circ G_\theta]$ takes a $(\kappa, \beta, \alpha, \eta)$ form. Then, by the sum rule for weak derivatives and the product rule of Lemma 2, $D^{\tilde{\alpha}_0}[\varphi \circ G_\theta] = D_{u_m}[D^{\alpha_0}[\varphi \circ G_\theta]]$ takes the form

$$D^{\tilde{\alpha}_0}[\varphi \circ G_\theta] = D_{u_m}[D^{\alpha_0}[\varphi \circ G_\theta]] = \sum_{i=1}^{I} \sum_{j=1}^{d^{\kappa_i}} D_{u_m}\left[D^{\beta_{ij}} \varphi \circ G_\theta\right] \prod_{l=1}^{\kappa_i} \partial^{\alpha_{ijl}} G_{\theta,\eta_{ijl}} + \sum_{i=1}^{I} \sum_{j=1}^{d^{\kappa_i}} \left[D^{\beta_{ij}} \varphi \circ G_\theta\right] \partial^{e[m]} \left[\prod_{l=1}^{\kappa_i} \partial^{\alpha_{ijl}} G_{\theta,\eta_{ijl}}\right]. \tag{12}$$

By the product rule for regular derivatives,

$$\partial^{e[m]}\left[\prod_{l=1}^{\kappa_i} \partial^{\alpha_{ijl}} G_{\theta,\eta_{ijl}}\right] = \sum_{l_0=1}^{\kappa_i} \partial^{\alpha_{ijl_0}+e[m]} G_{\theta,\eta_{ijl_0}} \prod_{\substack{l \in [1,\kappa_i] \\ l \neq l_0}} \partial^{\alpha_{ijl}} G_{\theta,\eta_{ijl}}. \tag{13}$$

Since $D^{\beta_{ij}}\varphi \in \mathcal{W}^{1,2}(\mathcal{X})$, the statement in Lemma 1 applies to its composition with G_θ, meaning

$$D_{u_m}\left[D^{\beta_{ij}} \varphi \circ G_\theta\right] = \sum_{j_0=1}^{d} \left[D_{x_{j_0}}\left[D^{\beta_{ij}} \varphi\right] \circ G_\theta\right] \frac{\partial G_{\theta,j_0}}{\partial u_m} = \sum_{j_0=1}^{d} \left[D^{\beta_{ij}+e[j_0]} \varphi \circ G_\theta\right] \frac{\partial G_{\theta,j_0}}{\partial u_m},$$

where, recall, $e[j_0]$ is a d-dimensional unit vector with 1 as the j_0'th element. Substituting these into (12), we get

$$D^{\tilde{\alpha}_0}[\varphi \circ G_\theta] = \sum_{i=1}^{I} \sum_{j=1}^{d^{\kappa_i}} \sum_{j_0=1}^{d} \left[D^{\beta_{ij}+e[j_0]} \varphi \circ G_\theta\right] \frac{\partial G_{\theta,j_0}}{\partial u_m} \prod_{l=1}^{\kappa_i} \partial^{\alpha_{ijl}} G_{\theta,\eta_{ijl}} + \sum_{i=1}^{I} \sum_{l_0=1}^{\kappa_i} \sum_{j=1}^{d^{\kappa_i}} \left[D^{\beta_{ij}} \varphi \circ G_\theta\right] \partial^{\alpha_{ijl_0}+e[m]} G_{\theta,\eta_{ijl_0}} \prod_{\substack{l \in [1,\kappa_i] \\ l \neq l_0}} \partial^{\alpha_{ijl}} G_{\theta,\eta_{ijl}} \tag{14}$$

Now all that is left to do is find $\tilde{I}, \tilde{\kappa}, \tilde{\beta}, \tilde{\alpha}, \tilde{\eta}$ for which this will that the $(\tilde{\kappa}, \tilde{\beta}, \tilde{\alpha}, \tilde{\eta})$–form similar to Equation (10). One can already see this should be possible, due to the flexibility in the definition of $(\tilde{\kappa}, \tilde{\beta}, \tilde{\alpha}, \tilde{\eta})$–forms; for completeness, we give the exact values now.

Define $\kappa_0 = 0$. Take $\tilde{I} = I + \sum_{i=1}^{I} \kappa_i$, $\tilde{\kappa}_i = \kappa_i + 1$ for $i \in [1, I]$ and $\tilde{\kappa}_i = \kappa_p$ when $i \in [I + \sum_{j=0}^{p-1} \kappa_j, I + \sum_{j=0}^{p} \kappa_j]$ for $p \in [1, I]$, $p \in \mathbb{N}$, and

$$\tilde{\beta}_{ij} = \begin{cases} \beta_{i\lfloor j/d \rfloor} + e[j \mod d], & i \in [1, I],\ j \in [1, 2^{\kappa_i+1}], \\ \beta_{pj}, & i \in (I + \sum_{j=0}^{p-1} \kappa_j, I + \sum_{j=0}^{p} \kappa_j],\ j \in [1, 2^{\kappa_p}] \text{ for } p \in [1, I], \end{cases}$$

$$\tilde{\alpha}_{ijl} = \begin{cases} \alpha_{i\lfloor j/d \rfloor l}, & i \in [1, I],\ j \in [1, 2^{\kappa_i+1}],\ l \in [1, \kappa_i], \\ e[m], & i \in [1, I],\ j \in [1, 2^{\kappa_i+1}],\ l = \kappa_i + 1, \\ \alpha_{pjl}, & i \in (I + \sum_{j=0}^{p-1} \kappa_j, I + \sum_{j=0}^{p} \kappa_j],\ j \in [1, 2^{\kappa_p}],\ l \in [1, \kappa_p] \setminus \{i - I - \sum_{j=0}^{p-1} \kappa_j\} \text{ for } p \in [1, I], \\ \alpha_{pjl} + e[m], & i \in (I + \sum_{j=0}^{p-1} \kappa_j, I + \sum_{j=0}^{p} \kappa_j],\ j \in [1, 2^{\kappa_p}],\ l = i - I - \sum_{j=0}^{p-1} \kappa_j \text{ for } p \in [1, I], \end{cases}$$

$$\tilde{\eta}_{ijl} = \begin{cases} \eta_{i\lfloor j/d \rfloor l}, & i \in [1, I],\ j \in [1, 2^{\kappa_i+1}],\ l \in [1, \kappa_i], \\ j \mod d, & i \in [1, I],\ j \in [1, 2^{\kappa_i+1}],\ l = \kappa_i + 1, \\ \eta_{pjl}, & i \in (I + \sum_{j=0}^{p-1} \kappa_j, I + \sum_{j=0}^{p} \kappa_j],\ j \in [1, 2^{\kappa_p}],\ l \in [1, \kappa_p] \text{ for } p \in [1, I]. \end{cases}$$

where $j \mod d$ is the remainder of dividing j by d. Then, (14) becomes

$$D^{\tilde{\alpha}_0}[\varphi \circ G_\theta] = \sum_{i=1}^{\tilde{I}} \sum_{j=1}^{d^{\tilde{\kappa}_i}} \left[D^{\tilde{\beta}_{ij}} \varphi \circ G_\theta\right] \prod_{l=1}^{\tilde{\kappa}_i} \partial^{\tilde{\alpha}_{ijl}} G_{\theta,\tilde{\eta}_{ijl}}.$$

This completes the proof of the induction step, and the theorem. □

A.3. Proof of Theorem 4

Before proving the main theorem, we introduce two auxilliary lemmas, Lemmas 3 and 4. The former will allow us to apply the chain rule of Theorem 6 to get $k(x, \cdot) \circ G_\theta \in \mathcal{H}_c$, and the latter claim the asymptotic rate of $m^{-\nu_c/s-1/2}$. The proof of Theorem 4 will follow.

Given A1 to A3, all that is missing to prove $k(x, \cdot) \circ G_\theta \in \mathcal{H}_c$ by applying Theorem 6 is the connection between RKHS of Matérn kernels, and Sobolev spaces. To that end, we introduce a Lemma (that is a minor extension to classic results, see for instance Wendland (2005, Corollary 10.48)) that links RKHS $\mathcal{H}_k$ of a Matérn kernel k of order ν with Sobolev spaces $\mathcal{W}^{l,2}$ for $l \in \mathbb{N}_+$. In the Lemma, we briefly refer to Bessel potential spaces—only for their norm-equivalence both to the RKHS of Matérn kernels, and to the Sobolev spaces of fractional order, which themselves lie between Sobolev spaces of integer order—and to extension operators, that allow us to extend results on $\mathbb{R}^d$ to open, connected, bounded $\mathcal{X}$ with a Lipschitz boundary. Every open, bounded, and convex $\mathcal{X}$ has a Lipschitz boundary (Stein, 1970); for example, this includes the hypercube $\mathcal{X} = (0, 1)^d$. For a detailed overview of Bessel potential spaces, fractional Sobolev spaces, and extension operators, we refer to Adams & Fournier (2003); these will only appear in the proof of the following Lemma.

For a $\beta \in \mathbb{R}_+ \cup \{0\}$, we denote the ceiling operation $\lceil \beta \rceil = \min(\{z \in \mathbb{N} \mid z \geq \beta\})$, and the rounding operation $\lfloor \beta \rfloor = \max(\{z \in \mathbb{N} \mid z \leq \beta\})$.

Lemma 3. *Suppose $\mathcal{X} = \mathbb{R}^d$, or $\mathcal{X} \subseteq \mathbb{R}^d$ is open, connected, and bounded with a Lipshitz boundary, and k is a Matérn kernel on $\mathcal{X}$ of order ν. Then, the RKHS $\mathcal{H}_k$ induced by k lies between Sobolev spaces $\mathcal{W}^{\lceil \nu+d/2 \rceil,2}(\mathcal{X})$ and $\mathcal{W}^{\lfloor \nu+d/2 \rfloor,2}(\mathcal{X})$, meaning*

$$\mathcal{W}^{\lceil \nu+d/2 \rceil,2}(\mathcal{X}) \subseteq \mathcal{H}_k \subseteq \mathcal{W}^{\lfloor \nu+d/2 \rfloor,2}(\mathcal{X}).$$

Proof. We start by proving the result for $\mathcal{X} = \mathbb{R}^d$. By Wendland (2005, Corollary 10.13), the RKHS of a Matérn k is norm-equivalent to the Bessel potential space $H^s(\mathcal{X})$ for $s = \nu + d/2$. The Bessel potential space $H^s(\mathcal{X})$, by Adams & Fournier (2003, Section 7.62), is norm-equivalent to a fractional Sobolev space (a Sobolev-Slobodeckij space) $W^{s,2}(\mathcal{X})$, which lies between spaces of integer order, $W^{\lceil s \rceil,2}(\mathcal{X}) \subseteq W^{s,2}(\mathcal{X}) \subseteq W^{\lfloor s \rfloor,2}(\mathcal{X})$.

Finally, the result $\mathcal{W}^{\lceil s \rceil,2}(\mathcal{X}) \subseteq \mathcal{H}_k \subseteq \mathcal{W}^{\lfloor s \rfloor,2}(\mathcal{X})$ extends to an open connected bounded $\mathcal{X} \subset \mathbb{R}^d$ with a Lipshitz boundary identically to the proof of Wendland (2005, Corollary 10.48), which makes use of the extension operator introduced for such $\mathcal{X}$ by Stein (1970). □

To show the claimed asymptotic rate, we use the following straightforward corollary of Wynne et al. (2021, Theorem 9).

Lemma 4 (Corollary of Theorem 9 in Wynne et al. (2021))**.** *Suppose for any $m \geq M \in \mathbb{N}_+$,*

- *$\mathbb{U}$ is a measure on a convex, open, and bounded $\mathcal{U} \subset \mathbb{R}^s$ that has a density $f_\mathbb{U} : \mathcal{U} \to [0, C'_\mathbb{U}]$ for some $C'_\mathbb{U} > 0$.*
- *$\{u_i\}_{i=1}^m$ are such that the fill distance $h_m = \mathcal{O}(m^{-1/s})$.*
- *$\{w_i\}_{i=1}^m$ are the optimal weights obtained based on the kernel c_{β_m} and measure $\mathbb{U}$, parametrised by $\beta_m \in B$ for some parameter space B,*
- *for any $\beta \in B$, c_β is a Matérn kernel of order ν_c; ν_c is independent of β.*

Then, for some C_0 independent of m and f, and any $f \in \mathcal{H}_c$ with $\|f\|_{\mathcal{H}_c} = 1$,

$$\left| \int_\mathcal{U} f(u)\mathbb{U}(du) - \sum_{i=1}^m w_i f(u_i) \right| \leq C_0 m^{-\nu_c/s-1/2}.$$

Proof. The expression on the left hand side of Wynne et al. (2021, Theorem 9) is $|\int_\mathcal{U} f(u)\mathbb{U}(\mathrm{d}u) - \sum_{i=1}^m w_i f(u_i)|$; the notation from their paper to this result maps as $\theta \to \beta$, $p \to f_\mathbb{U}$, $\mathcal{X} \to \mathcal{U}$, $x \to u$, $\Theta \to B$, and the prior mean $\mu(\beta) = 0$ for any $\beta \in B$. First, we show the assumptions in the Theorem hold.

Assumption 1 (Assumptions on the Domain): An open, bounded, and convex $\mathcal{U}$ satisfies the assumption, as discussed in Wynne et al. (2021).

Assumption 2 (Assumptions on the Kernel Parameters): Since c_β is a Matérn kernel of order ν_c, the smoothness constant of c_β is $\nu_c + s/2$ regardless of the value of $\beta \in B$, meaning $\tau(\beta) = \tau_c^- = \tau_c^+ = \nu_c + s/2 > s/2$. Lastly, the norm equivalence constants of Wynne et al. (2021, Equation 3) are the same for all β—since the respective RKHS and Sobolev spaces are—so the set of extreme values B_m^* is finite and does not depend on m; we denote $B_c^* = B_m^*$, to highlight that B_c^* only depends on the choice of kernel family c and not m.

Assumption 3 (Assumptions on the Kernel Smoothness Range): As discussed in Assumption 2, $\tau(\beta) = \nu_c + s/2$ for any $\beta \in B$, so the set in the statement of Assumption 3 has only one element.

Assumption 4 (Assumptions on the Target Function and Mean Function): The target function f is in $\mathcal{H}_c$, meaning $\tau_f = \tau_c^- = \tau_c^+ = \nu_c + s/2$. The mean function $\mu(\beta)$ was taken to be zero, so has zero norm.

Lastly, take h_0 such that $h_1 \leq h_0$; as we assumed $h_m = \mathcal{O}(m^{-1/s})$, it holds that $h_0 \leq h_m$ for all $m \geq 1$. Therefore, all the assumptions are satisfied and Wynne et al. (2021, Theorem 9) applies; moreover, the bounding expression is $C_0 m^{-\alpha/s}$ for $\alpha = \nu_c + s/2$ and some C_0 independent of m and f since

- $h_m = \mathcal{O}(m^{-1/s})$, and as $\tau_f = \tau_c^- = \tau_c^+ = \nu_c + s/2$ as discussed in the verification of assumptions, $h_m^{\max(\tau_f, \tau_c^-)} = O(m^{-\nu_c/s - 1/2})$,
- the rest of the multipliers do not depend on m and f: C depends only on $\mathcal{U}$, s, $\tau_f = \nu_c + s/2$, and B^*; $\|f_{\mathbb{U}}\|_{\mathcal{L}^2(\mathcal{U})}$ is a constant and finite since $f_{\mathbb{U}}$ is bounded above; $\tau_f - \tau_c^+ = 0$ so rising to its power produces 1; the norm $\|f\|_{\mathcal{H}_c} = 1$; for any $m \geq M$, $\mu(\beta_m) = 0$.

This completes the proof. □

Now we are ready to prove the main theorem.

Proof of Theorem 4. To show $k(x, \cdot) \circ G_\theta \in \mathcal{H}_c$ for all $x \in \mathcal{X}$, first note that Lemma 3 applies for both $\mathcal{U}$ and $\mathcal{X}$ that satisfy A1: trivially for $\mathbb{R}^d$, and for an open, convex, and bounded space since it has a Lipschitz boundary (Stein, 1970). Since by A3, k is a Matérn kernel of order ν_k, it holds by Lemma 3 that $k(x, \cdot) \in \mathcal{W}^{l_\varphi, 2}(\mathcal{X})$ for $l_\varphi = \lfloor \nu_k + d/2 \rfloor$ and any $x \in \mathcal{X}$. Then, by Theorem 6, $k(x, \cdot) \circ G_\theta \in \mathcal{W}^{\tilde{l}, 2}(\mathcal{U})$, for a G_θ that satisfies A2, and $\tilde{l} = \min(l_\varphi, l) = \min(\lfloor \nu_k + d/2 \rfloor, l)$. By A3, $\nu_c \leq \min(\lfloor \nu_k + d/2 \rfloor, l) - s/2 = \tilde{l} - s/2$, and it holds that $\tilde{l} \geq \nu_c + s/2$. Since $\tilde{l}$ is an integer, this implies $\tilde{l} \geq \lceil \nu_c + s/2 \rceil$, and we have that $\mathcal{W}^{\tilde{l}, 2}(\mathcal{U}) \subseteq \mathcal{W}^{\lceil \nu_c + s/2 \rceil, 2}(\mathcal{U})$. Finally, as c is a Matérn kernel or order ν_c, by Lemma 3 it holds that $\mathcal{W}^{\lceil \nu_c + s/2 \rceil, 2}(\mathcal{U}) \subseteq \mathcal{H}_c$, and we arrive at $k(x, \cdot) \circ G_\theta \in \mathcal{H}_c$.

Since $k(x, \cdot) \circ G_\theta \in \mathcal{H}_c$ holds, we can use Theorem 3 and state

$$|\mathrm{MMD}_k(\mathbb{P}_\theta, \mathbb{Q}^n) - \mathrm{MMD}_k(\mathbb{P}_\theta^m, \mathbb{Q}^n)| \leq K \times \mathrm{MMD}_c\left(\mathbb{U}, \textstyle\sum_{i=1}^m w_i \delta_{u_i}\right).$$

By the reproducing property, it holds that $\sup_{\substack{f \in \mathcal{H}_c \\ \|f\|_{\mathcal{H}_c} = 1}} |\int_{\mathcal{U}} f(u)\mathbb{P}(du) - \int_{\mathcal{U}} f(u)\mathbb{Q}(du)|$ is equal to $\mathrm{MMD}_c(\mathbb{P}, \mathbb{Q})$ for any two distributions $\mathbb{P}, \mathbb{Q}$ on $\mathcal{U}$. Then,

$$\mathrm{MMD}_c(\mathbb{U}, \textstyle\sum_{i=1}^m w_i \delta_{u_i}) = \sup_{\substack{f \in \mathcal{H}_c \\ \|f\|_{\mathcal{H}_c} = 1}} \left| \int_{\mathcal{U}} f(u)\mathbb{U}(du) - \sum_{i=1}^m w_i f(u_i) \right|.$$

The expression under the supremum is bounded by Lemma 4 with $C_0 m^{-\nu_c/s - 1/2}$, for C_0 independent of m and f. Therefore, $\mathrm{MMD}_c(\mathbb{U}, \sum_{i=1}^m w_i \delta_{u_i}) \leq C_0 m^{-\nu_c/s - 1/2}$, and the result holds. □

Note that while the result was formulated for the special case of convex spaces, it applies more generally to any open, connected, bounded $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{U} \subset \mathbb{R}^s$ with Lipschitz-continuous boundaries—with no changes to the proof. The applicability to $\mathcal{X} = \mathbb{R}^d$ remains unchanged; $\mathcal{U}$, however, must remain bounded for Theorem 6 to hold.

A.4. Computational and sample complexity

We derive the condition under which the OW estimator achieves better sample complexity than the V-statistic for the same order of computational cost, see Table 3 for the rates.

Suppose the cost for both V-statistic and OW is $\mathcal{O}(\tilde{m})$. Then, the sample complexity for the V-statistic can be written in terms of $\tilde{m}$ as $\mathcal{O}(\tilde{m}^{-1/4})$. Similarly, for the OW estimator, the sample complexity in terms of $\tilde{m}$ is $\mathcal{O}(\tilde{m}^{-(\nu_c + \frac{s}{2})/3s})$. The more accurate estimator is therefore the one whose error rate goes to zero quicker. Therefore, the OW estimator is more accurate than the V-statistic if

$$\tfrac{\nu_c}{s} > \tfrac{1}{4},$$

which for the common choice of $\nu_c = 5/2$ implies $s < 10$.

Table 3. Computational and sample complexity rates of the V-statistic and the OW estimator with respect to m.

	Cost	**Error**
V-statistic	$\mathcal{O}(m^2)$	$\mathcal{O}(m^{-\frac{1}{2}})$
OW	$\mathcal{O}(m^3)$	$\mathcal{O}(m^{-\frac{\nu_c}{s}-\frac{1}{2}})$

A.5. Derivation of closed-form kernel embeddings

We have $z_i = \int_{\mathcal{U}} c(u_i, v)\mathbb{U}(\mathrm{d}v)$, where $c : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$ is the SE kernel parameterised by the lengthscale $l > 0$, i.e., $c(u, v) = \sqrt{2\pi} l \varphi(u; v, l^2)$, where φ is the Gaussian pdf. For $s > 1$, we can write the kernel as $c(u, v) = \prod_{j=1}^{s} c(u_j, v_j)$. We now derive closed-form kernel embeddings for z_i for different choices of the base space $\mathcal{U}$ and the distribution $\mathbb{U}$.

For $\mathcal{U} = [0, 1]^s$, and $\mathbb{U}$ the uniform distribution, i.e., $u_i \sim \text{Uniform}([0, 1]^s)$, we get

$$z_i = \prod_{j=1}^{s} \int_{[0,1]} c(u_{ij}, v_j)\mathrm{d}v_j = \prod_{j=1}^{s} \sqrt{2\pi} l \left[\varphi(1; u_{ij}, l^2) - \varphi(0; u_{ij}, l^2)\right],$$

where φ is the Gaussian cdf and u_{ij} is the jth element of u_i.

In the case of $\mathcal{U} = \mathbb{R}^s$, and $\mathbb{U}$ being the Gaussian distribution such that $u_i \sim \mathcal{N}(\mu, \Sigma)$, where $\mu = [\mu_1, \ldots, \mu_s]^\top$ and Σ is the s-dimensional diagonal matrix with entries $(\sigma_1^2, \ldots, \sigma_s^2)$, the closed-form embedding for z_i reads

$$\begin{aligned} z_i = \prod_{j=1}^{s} \int_{-\infty}^{\infty} c(u_{ij}, v_j)\varphi(v_j; \mu_j, \sigma_j^2)\mathrm{d}v_j &= \prod_{j=1}^{s} \sqrt{2\pi} l \int_{-\infty}^{\infty} \varphi(v_j; u_{ij}, l^2)\varphi(v_j; \mu_j, \sigma_j^2)\mathrm{d}v_j \\ &= \prod_{j=1}^{s} \sqrt{\frac{l^2}{(l^2+\sigma_j^2)}} \exp\left(\frac{-(u_{ij}-\mu_j)^2}{2(l^2+\sigma_j^2)}\right). \end{aligned}$$

For the special case of $\Sigma = \mathrm{diag}(\sigma^2, \ldots, \sigma^2)$, the expression simplifies to

$$z_i = \left(\frac{l^2}{l^2+\sigma^2}\right)^{s/2} \exp\left(-\frac{\|u_i-\mu\|^2}{2(l^2+\sigma^2)}\right).$$

B. Additional Experimental details

True parameter values of the benchmark simulators in Section 5.1 is given in Appendix B.1. Appendix B.2 and Appendix B.4 provide additional results and details regarding the experiments in Section 5.2. Finally, the link to the source code of the wind farm simulator is in Appendix B.5.

B.1. Benchmark Simulators

We now provide further details on the benchmark simulators. For drawing iid or RQMC points, we use the implementation from `SciPy` (Virtanen et al., 2020). Below, we report the parameter value θ used to generate the results in Table 1 for each model. We refer the reader to the respective reference in Table 1 for a description of the model and their parameters.

g-and-k distribution: $(A, B, g, k) = (3, 1, 0.1, 0.1)$

Two moons: $(\theta_1, \theta_2) = (0, 0)$

Bivariate Beta: $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5) = (1, 1, 1, 1, 1)$

Moving average (MA) 2: $(\theta_1, \theta_2) = (0.6, 0.2)$

M/G/1 queue: $(\theta_1, \theta_2, \theta_3) = (1, 5, 0.2)$

Lotka-Volterra: $(\theta_{11}, \theta_{12}, \theta_{13}) = (5, 0.025, 6)$

B.2. Multivariate g-and-k

The performance of the V-statistic and our OW estimator as a function of θ_3 parameter of the multivariate g-and-k distribution is shown in Figure 7 (left). The observed effect on the performance is similar to that of Figure 3c, where the error in the OW estimator increases as we vary θ_3. The degradation in performance is not as severe as when varying θ_4, indicating that the smoothness of the multivariate g-and-k generator is not impacted by θ_3 compared to θ_4. Both the uniform and the Gaussian embedding achieves better performance than the V-statistic, whose performance remains unaffected by θ_3.

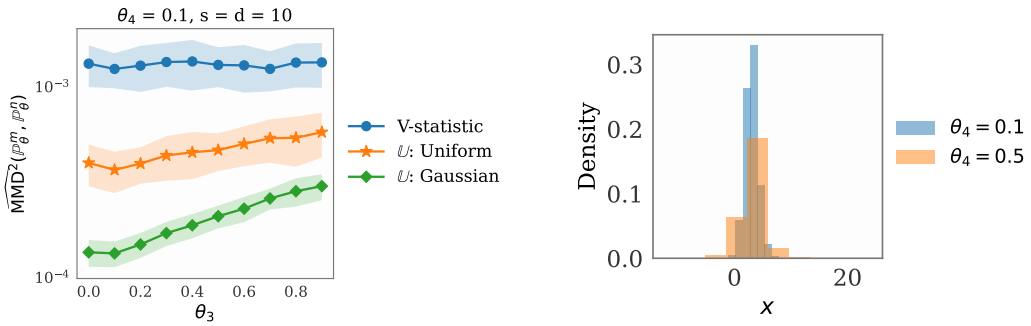


Figure 7. Additional results for the multivariate g-and-k distributions. *Left:* Estimated MMD^2 for the V-statistic and our OW estimator as a function of θ_3. *Right:* Histogram of samples from the g-and-k distribution for different values of θ_4. Settings: no. of samples = 100,000, $\theta_1 = 3$, $\theta_2 = 1$, $\theta_3 = 0.1$.

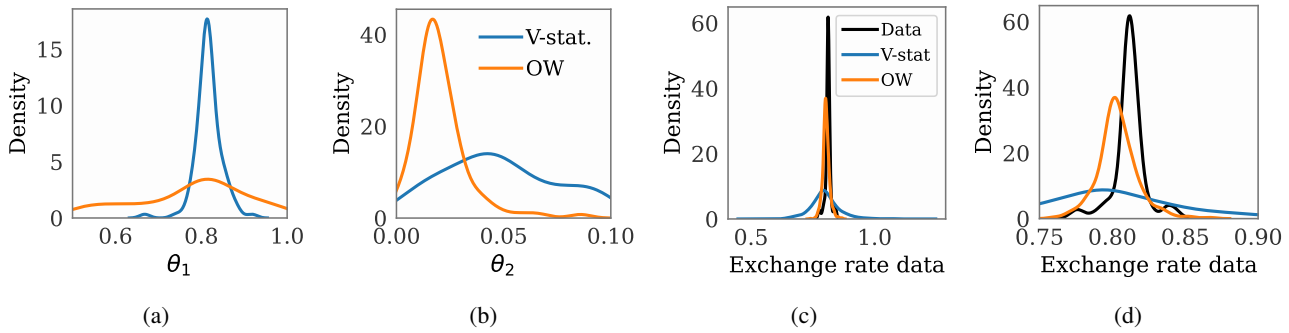


Figure 8. US dollar/Canadian dollar exchange rate data experiment on the g-and-k distribution. (a) ABC posterior for θ_1. (b) ABC posterior for θ_2. (c) Model fit based on the MAP estimates of θ_1 and θ_2 for V-statistic (in blue) and OW estimator (in orange). The OW estimator leads to a much better fit to the exchange rate data (shown in black). (d) Zoomed-in version of (c).

B.3. Exchange rate data experiment

We apply the g-and-k simulator to the US dollar to Canadian dollar exchange rate data (Verbeek, 2018) from the `Ecdat` R package. The data is shown in Figure 8c in black, which has $n = 501$ points. We fit the univariate g-and-k model to this data using the ABC method of Equation (4), with both the V-statistic and our OW estimator. For simplicity, we keep $\theta_3 = 0.12$ and $\theta_4 = 0.35$ fixed and only estimate the first two parameters. We set $m = 20$ and simulate 2000 parameter values from the prior $\mathcal{U}([0.5, 1] \times [0, 0.1])$. The resulting ABC posteriors with tolerance $\varepsilon = 5\%$ are shown in Figure 8a and Figure 8b for θ_1 and θ_2, respectively. Figure 8c shows the corresponding predictions based on the MAP estimate of the ABC posteriors. We observe that our OW estimator leads to a better fit than the V-statistic estimator. We are able to estimate the variance of the data (governed by θ_2) much more accurately than the V-statistic, which overestimates the variance.

B.4. Composite goodness-of-fit test: details and additional results

Algorithm 1 shows the details of the composite goodness-of-fit test using the parametric bootstrap. The algorithm is written for the V-statistic estimator, but each instance of the squared MMD can be replaced with our OW estimator. In practice, to compute $\arg\min_\theta \text{MMD}^2(\mathbb{P}_\theta, \mathbb{Q}^n)$ we use gradient-based optimisation, as described in Algorithm 2. The definitions of the hyperparameters of these two algorithms, and the values that we use, are given in Table 4.

Θ_{init} is the distribution from which the initial parameters are sampled, and is a uniform distribution with the following ranges: $\theta_1 : (0.001, 5)$, $\theta_2 : (0.001, 5)$, $\theta_3 : (0.001, 1)$, $\theta_5 : (0.001, 1)$. To compute the fraction of times that the null hypothesis is rejected (Table 2) we repeat the experiment 150 times.

Algorithm 1: Composite goodness-of-fit test

Input: $\mathbb{P}_\theta$, $\mathbb{Q}^n$, α, B
$\hat{\theta}_n = \arg\min_\theta \mathrm{MMD}^2(\mathbb{P}_\theta, \mathbb{Q}^n)$;
for $k \in \{1, \ldots, B\}$ **do**
  $\mathbb{Q}^n_{(k)} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i^{(k)}}$, $\{x_i^{(k)}\}_{i=1}^n \sim \mathbb{P}_{\hat{\theta}_n}$;
  $\hat{\theta}^n_{(k)} = \arg\min_{\theta\in\Theta} \mathrm{MMD}^2(\mathbb{P}_\theta, \mathbb{Q}^n_{(k)})$;
  $\Delta_{(k)} = \mathrm{MMD}^2(\mathbb{P}_{\hat{\theta}^n_{(k)}}, \mathbb{Q}^n_{(k)})$;
$c_\alpha = \mathrm{quantile}(\{\Delta_{(1)}, \ldots, \Delta_{(B)}\}, 1-\alpha)$;
$\mathbb{P}^m_{\hat{\theta}_n} = \frac{1}{m}\sum_{i=1}^m \delta_{y_i}$, where $\{y_i\}_{i=1}^m \sim \mathbb{P}_{\hat{\theta}_n}$;
if $\mathrm{MMD}^2(\mathbb{P}^m_{\hat{\theta}_n}, \mathbb{Q}^n) > c_\alpha$ **then**
  return reject;
else
  return do not reject;

Algorithm 2: Random-restart optimiser

Input: $\mathbb{P}_\theta$, $\mathbb{Q}^n$, m, I, R, S, s, Θ^{init}
Function *loss(θ)* **is**
  $\mathbb{P}^m_\theta = \frac{1}{m}\sum_{i=1}^m \delta_{y_i}$, where $\{y_i\}_{i=1}^m \sim \mathbb{P}_\theta$;
  return $\mathrm{MMD}^2(\mathbb{P}^m_\theta, \mathbb{Q}^n)$;
$\theta^{\text{trial}}_{(1)}, \ldots, \theta^{\text{trial}}_{(I)} \sim \Theta^{\text{init}}$;
Select $\theta^{\text{init}}_{(1)}, \ldots, \theta^{\text{init}}_{(R)} \in \{\theta^{\text{trial}}_{(k)}\}_{k=1}^I$ that yield the smallest $\mathrm{loss}(\theta^{\text{init}}_{(k)})$;
$\hat{\theta}^{\text{opt}}_{(1)}, \ldots, \hat{\theta}^{\text{opt}}_{(R)} =$ **for** $k \in \{1, \ldots, R\}$ **do**
  $\hat{\theta}^{\text{opt}}_{(k)} = \text{adam_optimizer}(\text{loss}, S, s, \theta^{\text{init}}_{(k)})$
return $\theta^* \in \{\hat{\theta}^{\text{opt}}_{(k)}\}_{k=1}^R$ s.t. $\forall k. \mathrm{loss}(\theta^*) \leq \mathrm{loss}(\hat{\theta}^{\text{opt}}_{(k)})$;

Table 4. Definitions of the hyperparameters.

hyperparameter	value	
α	0.05	level of the test
B	200	number of bootstrap samples
m	100	number of samples from the simulator
n	500	number of observations in the data
I	50	number of initial parameters sampled
R	10	number of initial parameters to optimise
S	200	number of gradient steps
s	0.04	step size

B.5. Large scale wind farm model

We include the comparison with the U-statistic estimator of MMD for the wind farm experiment in Figure 9. Observations are similar to that of Figure 5 — our OW estimator leads to much more concentrated ABC posteriors around the true value than the U-statistic.

The low-order wake model is described in Kirby et al. (2023) and the code is available at https://github.com/AndrewKirby2/ctstar_statistical_model/blob/main/low_order_wake_model.py.

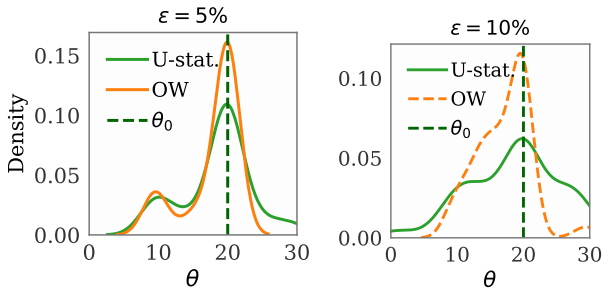


Figure 9. ABC posteriors for the wind farm model. Our OW estimator yields posterior samples that are more concentrated around the true θ_0 than the U-statistic. Settings: $n = 100$, $\theta_0 = 20$.